The landscape epidemiology of canine rabies virus in Tanzania

by

Kirstyn Brunker

Submitted in fulfilment of the requirements for the

Degree of *Doctor of Philosophy* (Ph.D.)

Institute of Biodiversity, Animal Health & Comparative Medicine

College of Medical, Veterinary and Life Sciences

University of Glasgow

May 11, 2016

Abstract

Infectious diseases pose a significant threat to animal and human health across the globe, with much of the burden falling on low-income countries. Despite efforts to control many of these diseases, very few have ever been eradicated. Their dynamics are often embedded in complex, heterogeneous landscapes defined by interacting population and landscape level processes. As such, landscape heterogeneity plays a key role in driving disease transmission and persistence. Incorporating landscape heterogeneity in studies of pathogen dynamics is challenging but the accessibility of data, particularly next generation sequencing data, has opened new avenues of research. Landscape epidemiology involves using an integrated approach to understand spatial patterns of disease, using methods that combine landscape genetics, ecology and epidemiology. In this thesis I use these integrative methods to determine the underlying mechanisms facilitating the spread and persistence of canine rabies virus in Tanzania. Whole genome level characterisation of rabies virus samples was achieved and used in combination with cutting-edge inference techniques to explore spatial patterns of rabies at different spatial scales.

Phylogeographic patterns were able to characterise spatial scales of endemic rabies transmission in Tanzania, uncovering strong viral population structure at sub-continental levels with evidence of a more fluid dispersal dynamic at local ($<$100km^2 area) spatial scales . Within-country phylogeographic patterns revealed large regional movements within Tanzania that could be attributed to human-mediated movements and revealed the presence of multiple co-circulating lineages within a single administrative district.

Finely resolved incidence data from the Serengeti District complemented with whole genome sequences enabled the exploration of local scales of transmission in more detail. By extending phylogeographic diffusion models to incorporate landscape heterogeneity I was able to uncover evidence supporting landscape predictors of rabies diffusion. While much of the spatial structure was attributable to the effects of isolation by distance, landscape predictors had discernible effects on diffusion. In particular, rivers appeared to act as a barrier to dispersal and road networks facilitated diffusion and I found evidence to support vaccination as an effective control measure for canine rabies in the Serengeti District. Importantly, I also found evidence to support vaccination as resistance to diffusion and therefore an effective control measure for dog rabies.

As a complementary approach a space-time-genetic algorithm was used to determine who-infected-whom in the Serengeti District. The model explicitly accounted for the possibility of exogenous sources of infection and how to incorporate genetic data available for only a proportion of samples. Direct transmission events were estimated between 42% of observed cases and highlighted the co-circulation of two major lineages in both time and space. Direct transmission events predominantly occurred over very small distances, <1km, but a large proportion of cases had unobserved sources that could represent transmission from dogs in neighbouring regions or larger indirect transmission events. A future development of the model is to delineate between these possibilities to assess the true contribution of exogenous sources to the system dynamic.

Ultimately these integrative models are at an early stage of development but highlight the power of genetic data to delineate fine-scale transmission patterns. The results from this thesis suggest that landscape features such as rivers could be exploited as barriers in step-wise vaccination campaigns and highlight the utility of genetic surveillance to monitor control and elimination as rabies management progresses.

Contents

List of Figures

List of Tables

I dedicate this thesis to my grandparents Robert and Joan Esler.

Acknowledgements

My PhD would not have been possible without the help of a great number of people. First and foremost, I would like to thank the supervisor dream-team, Roman Biek and Katie Hampson, for their encouragement, support and guidance throughout my PhD. They have gone above and beyond to provide opportunities and instil confidence in my abilities as a scientist.

For the last four years I have had the pleasure to work in an institute full of a strange yet wonderful collective of people. Pub nights, coffee mornings, feral mice, random excursions and (sometimes) work-related endeavours have all contributed to an fantastic PhD experience! Special mention must go to my office mates "Wild Bill", Nardus, Hannah, Miriam, Sonia, and Laurie for a multitude of tea breaks, laughs and support that have kept me going through my PhD.

I must also thank the many collaborators that have made this project possible. Many thanks to the fantastic rabies group at APHA in Weybridge, especially Dan Horton and Denise Marston and to Richard Ellis for allowing me to use his lab and pester him for reagents.
My PhD was funded by an MRC studentship and I would like to thank MRC for providing this opportunity and for extra funding granted via an MRC supplement award, which facilitated me to travel and stay in Belgium for work in Philippe Lemey's group. I would like to thank Philippe and his group for welcoming me to Leuven and ensuring I was fully acquainted with a range of Belgian beer whilst simultaneously being schooled on phylodynamic inference.
A very big thank you to my collaborator Samuel Soubeyrand for his endless patience and help as I struggled to come to terms with the beauty of his space-time-genetic models.
Many thanks must go to our collaborators in Tanzania including the Ministries of Livestock and Fisheries Development and of Health and Social Welfare, Tanzania National Parks, Tanzania Wildlife Research Institute, Ngorongoro Conservation Area Authority, Tanzania Commission for Science and Technology, and National Institute for Medical Research for permission and collaboration; and the Arusha, Mtwara and Mwanza Veterinary Investigation Centres, Pemba Veterinary Office and the Frankfurt Zoological Society for logistical and technical support. I am especially grateful to the hard-working research assistants who have collected rabies samples during and beyond my PhD: Joel Changalucha, Zilpah Kaare, Gurdeep

Kour, Ahmed Lugelo, Kennedy Lushasi, Mathias Magoto, Khadija Said, and Renatus Herman.

In the realms outside the university I would like to thank my parents for putting up with my long-term student status and general moans about life, my little brother for just being my little brother and my grandparents for their encouragement and support. Many thanks to Fiona and the ginger dude for putting me up in London on the many times I didn't want to stay near a lab in Weybridge and for generally being entertaining in so many ways!

Lastly, particular thanks to Lyall, for not being a scientist and loving and supporting me through the ups and downs of my PhD.

Declaration

I declare that the work in this thesis is my own, except where otherwise stated. Much of the material included in this thesis has been produced in co-authorship with others and some has been presented for publication. My personal contribution to each chapter is as follows:

Chapter 2: *Published as*: Brunker, K., Hampson, K., Horton, D. L., & Biek, R. (2012). Integrating the landscape epidemiology and genetics of RNA viruses: rabies in domestic dogs as a model. *Parasitology*, 139(14), 1899-1913. doi: 10.1017/S003118201200090X.

Literature review and chapter drafted by KB. Final draft enhanced by KH, DH and RB.

Chapter 3: *Published as*: Brunker, K., Marston, D. A., Horton, D. L., Cleaveland, S., Fooks, A. R., Kazwala, R.,Ngeleja, C. Lembo, T., Sambo, M., Mtema, Z.J., Sikana, L., Wilkie, G.,Biek, R., Hampson, K.(2015). Elucidating the phylodynamics of endemic rabies virus in eastern Africa using whole-genome sequencing. *Virus Evolution*, 1(1), vev011. doi:10.1093/ve/vev011.

Molecular work and sequencing by KB, DM and GW. Data analysis and chapter drafted by KB and enhanced by KH and RB.

Chapter 4: *In preparation for submission as*: Brunker, K., Lemey, P., Hampson, K., Biek, R. Quantifying the effects of landscape heterogeneity on the local-scale phylodynamics of an endemic zoonotic virus.

Initial concept developed by PL and KB. Data analysis and chapter drafted by KB. Final draft enhanced by KH and RB.

Chapter 5: *In preparation for submission as*: Brunker, K., Soubeyrand, S., Hampson, K., Biek, R. Inferring the dynamics of endemic canine rabies virus using high resolution space-time-genetic data.

Initial concept developed by SS, KH and KB. Algorithm by SS. Data analysis and chapter drafted by KB, with method text adapted from SS. Appendix 2A by SS. Final draft enhanced by KH and RB.

I further declare that no part of this work has been submitted as part of any other degree.

Kirstyn Brunker

CHAPTER 1

Overview

The dynamics of infectious diseases are structured by contact between infectious and susceptible individuals, a process inherently determined by host abundance, distribution and movements. As hosts are embedded in a complex landscape of ecological processes it is not surprising that spatial heterogeneity plays a critical role in structuring transmission events and perpetuating the spread of disease. Ecological research has long had a fascination with the interplay between spatial patterns and ecological processes and there are many examples of how landscape attributes influences animal and plant populations. In turn, epidemiologists are keen to understand the ecological and evolutionary aspects of infectious diseases and the importance of spatial structuring in disease systems.

Landscape epidemiology is a field founded on the integration of approaches and concepts from landscape and disease ecology. The term was originally coined by the parasitologist Evgeniy Pavlovsky in the 1930s, who introduced the concept of natural nidality in human diseases (Pavlovsky & Levine, 1966). The field has regained momentum more recently (McCallum, 2008; Meentemeyer *et al.*, 2012; Ostfeld *et al.*, 2005) with the advent of next generation sequencing technologies and new methods for spatial analyses, which provide unprecedented access to pathogen genetic data and ways to explore spatial heterogeneity.

This thesis is concerned with the application of landscape epidemiological approaches to determine the mechanisms underlying the spread and persistence of infectious diseases. Specifically, I explore this using rabies virus as a model system. Rabies virus is a globally distributed multi-host zoonotic pathogen that is maintained in distinct host-species associated transmission cycles. Of most concern to human health is the persistence of rabies in domestic dog (*Canis familiaris*) populations, with the vast majority of (the tens of thousands of) human rabies cases caused by bites from rabid dogs. Despite the importance of domestic dogs as the principal reservoir of rabies virus in Asia and Africa, where most of these deaths occur, we still have little insight into the underlying mechanisms governing viral dynamics in this host. As a directly transmitted pathogen, patterns of rabies transmission are rooted in host dynamics influenced by landscape processes.

Phylogenetic inference is central to this thesis, providing the basis for increasingly sophisticated methods to study the processes influencing rabies spread at various spatial scales. Phylogenetic trees (phylogenies) represent evolutionary histories and relationships between individuals or groups of organisms (see Chapter 1 2.1 for basic visualisation and terminology) inferred from molecular sequence data. The incorporation of additional information in this phylogenetic framework e.g. temporal, host, phenotypic or geographic sampling data (a field known as phylodynamics), has become a popular means to extract key information on spatio-temporal patterns of pathogen spread and the interplay between evolutionary and epidemiological processes. A range of phylodynamic frameworks are explored though Chapters 3 and 4, complemented by an alternative but related approach involving transmission tree reconstructions in Chapter 5.

In Chapter 2 I review the integration of landscape genetic data into landscape epidemiology and how this can be used to answer key questions regarding the spread and persistence of disease, with a particular focus on RNA viruses. These viruses are particularly amenable to studies exploring the impact of landscape processes on the evolutionary dynamics of pathogens due to their high evolutionary rates, meaning their evolution can be tracked as a response to underlying epidemiological or ecological processes in short timeframes. Rabies virus belongs in this category of viruses and is the specific focus of my thesis.

The remainder of the thesis is split into chapters exploring the direct application of landscape genetic and spatial methods to explore the themes and questions identified in Chapter 2 in the context of spatial scale. The problems of pattern and scale are widely recognised in ecology (Levin, 1992) and uncovering the processes important in overall disease dynamics requires assessing multiple levels in the organisational scale of disease transmission.

In Chapter 3 a phylodynamic framework is used to explore the range of spatio-temporal patterns observable in African rabies virus populations from sub-continental to local scales and determine the landscape processes responsible. Whereas Chapters 4 and 5 focus down on spatial processes at the local endemic scale i.e. within a single administrative district. Here I invoke cutting-edge techniques to incorporate all available data in powerful Bayesian inference schemes to explore various aspects of pathogen dynamics at a scale that has not been attempted before for canine rabies. In Chapter 4 this involves ways to explicitly incorporate landscape heterogeneity in phylodynamic analyses in order to test and quantify significant predictors of viral diffusion. Identifying landscape drivers of diffusion is challenging but can provide information on the determinants of viral spread to inform strategies and interrupt the spatial spread of disease (Lemey *et al.*, 2014; Nelson *et al.*, 2015).

In Chapter 5 I focus on the foundation of infectious disease dynamics by reconstructing who-infected-whom. Transmission is the most critical yet elusive event in infectious disease epidemiology but we know little about it. Fine-grained viral genetic information and spatially resolved incidence data can provide important insights on the transmission process, yet combining these data remains a major statistical challenge. I present a framework to synergistically integrate epidemiological and genetic data to trace transmission trees and enhance our understanding of the transmission processes of endemically circulating rabies.

Overall, I demonstrate how viral genome sequence data can be used to answer key research questions regarding the landscape and evolutionary processes that give rise to spatial patterns of disease and how this information can he used to enhance and inform infectious disease control and public health policy (Chapter 6).

CHAPTER 2

Integrating the landscape epidemiology and genetics of RNA viruses: rabies in domestic dogs as a model.

2.1 Abstract

Landscape epidemiology and landscape genetics combine advances in molecular techniques, spatial analyses, and epidemiological models to generate a more real-world understanding of infectious disease dynamics and provide powerful new tools for the study of RNA viruses. Using dog rabies as a model I have identified how key questions regarding viral spread and persistence can be addressed using a combination of these techniques. In contrast to wildlife rabies, investigations into the landscape epidemiology of domestic dog rabies requires more detailed assessment of the role of humans in disease spread, including the incorporation of anthropogenic landscape features, human movements and socio-cultural factors into spatial models. In particular, identifying and quantifying the influence of anthropogenic features on pathogen spread and measuring the permeability of dispersal barriers are important considerations for planning control strategies, and may differ according to cultural, social and geographical variation across countries or continents. Challenges for dog rabies research include the development of metapopulation models and transmission networks using genetic information to uncover potential source/sink dynamics and identify the main routes of viral dissemination. Information generated from a landscape genetics approach will facilitate spatially strategic control programmes that accommodate for heterogeneities in the landscape and therefore utilise resources in the most cost-effective way. This can include the efficient placement of vaccine barriers, surveillance points and adaptive management for large-scale control programmes.

2.2 Introduction

Landscape epidemiology is the study of the causes and consequences of spatial variation in disease incidence or risk across heterogeneous landscapes (Ostfeld *et al.*, 2005). Landscape structure affects the distribution, abundance and movements of host, vector and pathogen populations and therefore inherently influences localised interactions between infectious and susceptible individuals (McCallum, 2008; Ostfeld *et al.*, 2005). Revealing the landscape factors underlying these interactions calls for an interdisciplinary approach that draws on a range of techniques across different spatial scales (Manel *et al.*, 2003; Ostfeld *et al.*, 2005). Molecular markers provide a basis for this by genetically tracking spatial and temporal dynamics in pathogen and host populations (Biek & Real, 2010). A landscape genetics approach to infectious disease therefore encompasses a range of analytical tools, including geographic information systems, remote sensing, population genetics, phylogenetics and statistical and mathematical modelling techniques (Manel *et al.*, 2003).

RNA viruses represent an ideal group for exploring landscape influences on evolutionary trajectories due to their characteristically high mutation rates and short generation times, which

means that epidemiological and population genetic processes occur on a similar timescale (Drummond *et al.*, 2003). The accumulation of mutations over time and space imprints on the structure of viral genomes in a population and can be visualised in data collected over months or years, providing a valuable resource for the elucidation of ecological and evolutionary dynamics. Despite this, the processes that govern phylogeographic patterns in viruses are still poorly understood (Holmes & Grenfell, 2009), pointing to the need for more detailed study into the effect of spatial heterogeneity on viral transmission.

Modern sequence analysis has the power to reveal the historical emergence of pathogen variants, distribution patterns, spillover events, interspecies transmission and changing selection pressures (Anderson *et al.*, 2010; Archie *et al.*, 2009). Advances in sequencing technologies have paved the way for a new generation of approaches to the study of disease dynamics, and next generation sequencing (NGS) techniques are being continually refined, improving and accessibility (Holmes & Grenfell, 2009; Metzker, 2010). I envision a paradigm shift to whole genome sequencing as the standard technique for characterising RNA viral evolution on small spatio-temporal scales, providing greater discrimination between genotypes and finer resolution in population structure (Holmes & Grenfell, 2009). Given the progress in sequencing technology and parallel advances in spatial analytical tools, it is an exciting time to study the landscape epidemiology of RNA viruses from a population genetic perspective.

In this review, I focus on rabies virus (RABV), which presents an excellent model system to illustrate the challenges and prospects of such an approach. RABV is a single stranded, negative-sense RNA virus belonging to the *Lyssavirus* genus (Family: Rhabdoviridae) (Dietzschold *et al.*, 2005). It is globally distributed and has the ability to infect all mammals, but typically exists in endemic foci as a reservoir host-specific variant with occasional spillover to other species (Rupprecht *et al.*, 2002). Domestic dogs, *Canis familiaris*, are the principal reservoir of RABV, responsible for 99% of the estimated 55,000 human deaths due to rabies that occur mainly in Asia and Africa every year (Knobel *et al.*, 2005).

Rabies has proven a remarkably valuable system for exploring the effect of landscape processes on host/pathogen interactions. However, most studies have focused on wildlife rabies due to the higher quality of surveillance data and availability of resources in areas with major endemic wildlife foci, e.g. raccoon rabies in eastern North America or fox rabies in Western Europe (Biek *et al.*, 2007; Bourhy *et al.*, 1999; Holmes, 2004; Smith *et al.*, 2002; Szanto *et al.*, 2011; Wheeler & Waller, 2008). The wealth of research into wildlife rabies provides a basis for comparison with the domestic dog foci that exist in Africa, Asia and parts of Latin America, which, despite their much greater public health burden, have been less well studied.

Domestic dogs are inherently tied to human populations, and various aspects of human ecology, including distribution, habitation and movement patterns, or cultural practices, will directly influence rabies spread in dog populations. Settlements can be considered dog 'habitat', and dog densities have been predicted on the basis of human demographics and human

geography (Butler & Bingham, 2000; Knobel *et al.*, 2008). A seemingly ubiquitous feature of countries with persistent dog rabies foci is the free-roaming nature of these populations, often referred to as 'neighbourhood' dogs. Dogs in Africa, Asia and parts of Latin America, where canine rabies is endemic, are rarely restricted by leashing or enclosures and their role as domestic animals varies e.g. watch dogs, trade, companion animals. The free movement of dogs would thus be expected to contribute to local rabies transmission, potentially resembling the known features of wildlife rabies and illustrating the complex interplay between anthropogenic and natural drivers of disease spread in this system. A key challenge is to uncover the extent to which natural constraints to rabies flux hold in a host-pathogen system with greater human-mediated dispersal; and what affect this has on phylogeographic signatures.

This review aims to synthesise our current understanding of rabies landscape epidemiology derived from genetic data across different spatio-temporal scales (see Table 2.1 for a list of relevant studies and associated analytical methods). Using the extensive work on wildlife rabies as a backdrop, I seek to identify commonalities as well as fundamental differences characterising the dynamics of RABV in domestic dog populations. I will argue that the specific 'landscape' supporting the sustained transmission of the virus in dogs is determined by a complex mixture of physical and human geography and that a quantitative understanding of these landscapes will be essential for rabies control and eradication. In addition, control activities are themselves predicted to change molecular epidemiological trajectories, creating interesting opportunities for adaptive management in rabies. Throughout the review I highlight how novel tools and technologies are being used to tackle these problems and identify key areas in which such approaches have future potential (see Box 2.1 for a glossary of key words highlighted through the review and Box 2.2 for a list of key research questions for dog rabies).

Box 2.1: Glossary

Association index (AI): test statistic used as a means of quantifying phylogeny-trait associations i.e. given a discrete character for each tip of a phylogenetic tree, are more closely related taxa likely to share the same trait values than is expected by chance alone. Avoids the issue of lack of independence due to shared ancestry (Wang *et al.*, 2001). AI values range from 0 indicating complete population subdivision to 1 indicating complete panmixis. Example traits of interest: geographic location, host species, physical characteristics.
Bayesian: a branch of statistics that focuses on estimating the posterior probability of a hypothesis. Posterior probability is interpreted as the confidence that the hypothesis is correct given the data. This quantity can be estimated using Bayes' Theorem as the product of the prior probability and the likelihood of an event.
Clade: a grouping of biological taxa that includes a common ancestor and all the descendants of that ancestor.
Likelihood: the probability of the data given the hypothesis (in contrast to Bayesian interpretation of probability). Maximum likelihood estimation aims to find a point estimate for the parameters that maximise the likelihood.
Metapopulation: concept describing the persistence of a species in a spatially heterogeneous environment as a balance between colonisation and extinction in loosely coupled subpopulations or 'patches' with different within and between-patch dynamics, which can be applied to infectious diseases (Grenfell & Harwood, 1997).
Most recent common ancestor (MRCA): most recent individual from which all taxa in the group are directly descended.
Oral rabies vaccination (ORV): distribution of oral rabies vaccine baits as a strategy to control the spread of wildlife rabies. Used as a control measure for wildlife rabies in North America, and Europe.
Phylogeography: combines phylogenetics and biogeography to describe the contemporary pattern of an organism's geographic spread according to gene genealogies.
R_0 (basic reproductive number): the average number of secondary cases derived from a single infectious individual in an entirely susceptible population (Anderson & May, 1991).
Superspreader: an individual causing a disproportionately high number of secondary cases, compared to the mean (represented by R_0.), depicted in the long tail of a frequency distribution of secondary cases (Lloyd-Smith *et al.*, 2005).
Surfing mutation model: genetic variants at the front of an advancing wave of infection are swept to high frequencies during an epidemic peak, resulting in long-term dominance of colonising genetic lineages (Excoffier & Ray, 2008; Klopfstein *et al.*, 2006).

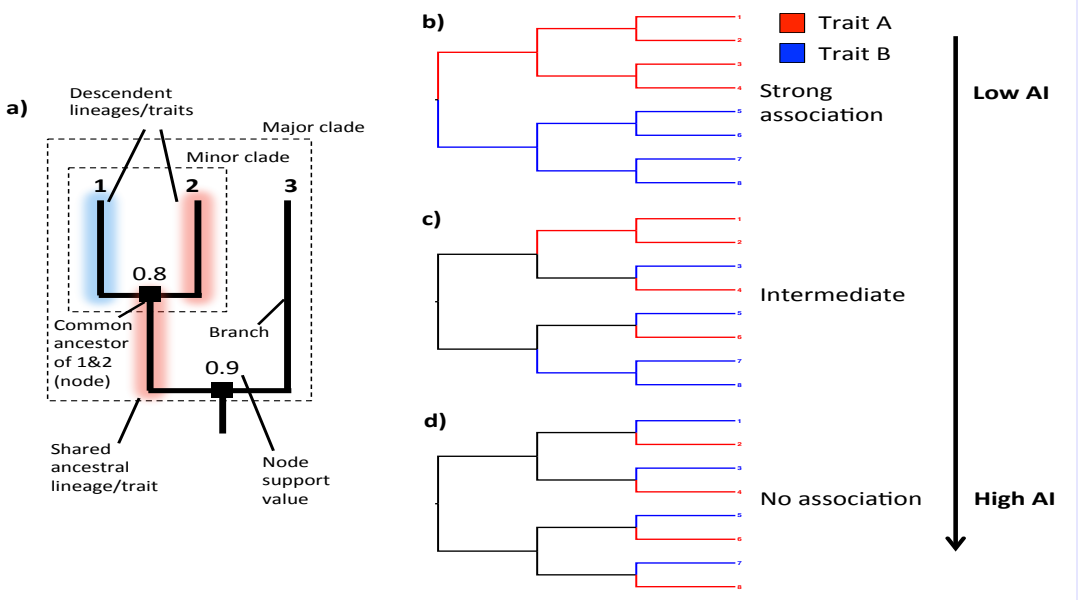


Figure I: Idealized phylogenies illustrating a) phylogenetic tree terminology, branch lengths represent a measure of chance such as time or nucleotide substitutions per site b) a strong phylogeny-trait relationship (low AI): tips with discrete character traits (red or blue) are tightly correlated with the phylogeny; c) intermediate scenario: appears to be some association of traits between sister taxa but also a level of admixture; d) no association (high AI): an unstructured phylogeny with no clear association between phenotype and phylogeny.

2.3 Rabies virus and transmission

Rabies virus has an enveloped, linear, non-segmented, negative-sense genome approximately 12 Kb in length with a relatively simple genome organisation. Genes encoding five structural proteins in the order nucleoprotein (N) - phosphoprotein (P) - matrix protein (M) - glycoprotein (G) - polymerase protein (L) are separated by non-coding intergenic regions (see Fig. 2.1). The genome is flanked by external signals at the 3' and 5' ends, which act as promotors for polymerisation and encapsidation respectively and are co-conserved for the first 11 positions at genome ends (Tordo *et al.*, 1988). Structurally the viral genome is embedded in monomers of the N protein and together with the P and L proteins forms a ribonucloprotein complex (RNP) involved in viral transcription and replication (Banerjee, 1987). The RNP associates with the M protein, which lies beneath the host-derived lipid membrane. G protein, the only external surface protein, protrudes from the lipid membrane to bind host cell receptors and is therefore considered important for pathogenesis (Dietzschold *et al.*, 2005). As an RNA virus RABV has a high rate of evolution (estimates of ~3.82 x 10^{-4} and ~3.25 x 10^{-4} for N and G genes respectively have previously reported for dog rabies (Talbi *et al.*, 2009)), which enables the accumulation of genetic variation on a monthly timescale (Biek *et al.*, 2015). Recombination appears to be rare in non-segmented negative-sense RNA viruses (Chare *et al.*, 2003; Han & Worobey, 2011), with little definitive evidence of it occurring in rabies virus populations (Chare *et al.*, 2003; Liu *et al.*, 2011). In order to facilitate recombination the same host cell would have to be co-infected with multiple viral strains. This scenario is limited by both the ecology of RABV, which is known to have geographically isolated viral strains (Biek *et al.*, 2007), and within host factors that may limit replication and contact between strains (Chare *et al.*, 2003).

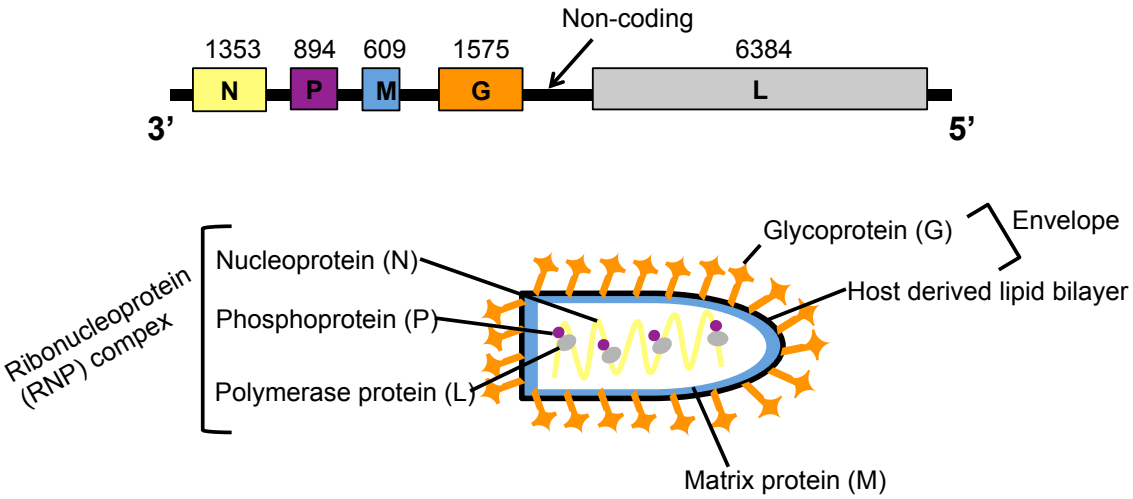


Figure 2.1: Rabies virus genome organisation and virion structure. The 12 Kb genome encodes five structural proteins in the order 3'-N-P-M-G-L-5' (gene lengths in base pairs indicated) with intergenic non-coding regions.

Transmission of rabies virus occurs predominantly through the bite of an infected animal, which inoculates virus-laden saliva into the subcutaneous and muscle tissue of a susceptible host (Dietzschold *et al.*, 2005). Once inoculated, the virus enters neurons and migrates to the central nervous system, before spreading to other organs, including the salivary glands where large amounts of infectious virions are shed into the saliva for further transmission (Dietzschold *et al.*, 2005). The incubation period is highly variable with a mean of 22 days in naturally infected dogs (Hampson *et al.*, 2009), but may extend to months or years due in part to localised replication (Hanlon *et al.*, 2007). In contrast, the infectious period is very short, around 3 days, and very rarely exceeds 10 days. The short infectious period severely restricts the spatial scale over which an infectious individual can transmit the virus, while the longer and more variable incubation time may permit the active or passive movement of infected individuals over larger distances, which, as detailed below, helps to explain *phylogeographic* patterns.

Box 2.2: Key research questions for dog rabies at the landscape scale

1. To what degree do the demography and ecology of dogs, as compared to humans, determine the dynamics of rabies transmission?
As dogs are inherently tied to humans, one might expect the patterns dictated by local dog movements to be confounded by human-mediated long-distance movements or anthropogenic features that facilitate connection of sub-populations.

2. Which aspects of human geography best predict landscape permeability to dog rabies? Are these predictors consistent across different areas and continents?
Although evidence suggests the influence of human geography on dog rabies dispersal, we have yet to uncover the best predictors of this form of spread. Specifically, a quantifiable method of characterising these landscape features would assist with the creation of guidelines for using dispersal barriers to aid control.

3. How does endemic rabies compare to epidemic spread?
Most studies have focused on outbreak situations in wildlife, but it is not clear how rabies dynamics change in systems where the pathogen has been circulating for centuries, as is the case for dog rabies. How long is phylogeographic structure maintained over time and do initial invasion pathways predict connectivity in the endemic state?

4. Does dog rabies persist in metapopulations?
Dog rabies may best be described as a series of sub-populations with varying inter- and intra-patch dynamics. Exploring the concept of a metapopulation dynamic for dog rabies presents a relatively unexplored area for future research and may uncover the mechanisms that allow pathogen persistence at a local scale.

5. How can information about landscape heterogeneity and genetic structure be incorporated into more efficient control programmes?
Information gathered from spatially explicit landscape models will allow targeted control measures using resources cost-effectively.Moreover, the implementation of vaccine barriers and other forms of control change the complexity of the landscape, potentially altering disease dynamics. Powerful genetic techniques may elucidate the effect of these new landscape heterogeneities, facilitating adaptive management.

2.4 Phylogenetic analysis

Phylogenetic analysis has played an important role in increasing our understanding of disease transmission (Holmes *et al.*, 1995; Pybus & Rambaut, 2009). Rapidly evolving pathogens, including many RNA viruses, generate detectable genetic differences over short observable time periods (Duffy *et al.*, 2008), to the effect that sequence data can be used to recover information on a pathogen population's evolutionary history and relationships (Pybus & Rambaut, 2009). Rabies virus belongs to this group of pathogens and as a consequence much of the methodology explored in this thesis is based on phylogenetic inference. Partial gene sequence is the most common publicly available data (usually the N gene) having been traditionally used for viral speciation and phylogenetic analysis (Marston *et al.*, 2013). However, progress in next generation sequencing technologies and the need for finer resolution characterisation has led to an increase in the production and availability of whole genome sequence data.

Phylogenetic trees (see Box 2.1 for example and terminology) reflecting evolutionary relationships are inferred based on a chosen tree-building methodology. While many tree-building methods are available (not described here, see (Holder & Lewis, 2003) for a review) they can be categorised according to whether a frequentist e.g. maximum *likelihood* (ML) or *Bayesian* approach is used to infer the phylogeny (see Box 2.3). Bayesian inference methods have become increasingly popular in phylogenetic analyses as they are less computationally intensive than ML and able to incorporate complex models of evolution (Huelsenbeck *et al.*, 2001). Furthermore, Bayesian inference can be effectively used in analyses based on the integration of genetic and epidemiological data (Pybus & Rambaut, 2009), which can quantitatively analyse the interaction between evolutionary and epidemiological processes in pathogen populations (a field termed phylodynamics (Grenfell *et al.*, 2004)). Such methods can implement sophisticated statistical models to infer the relationship between evolution and time (molecular clock models (Drummond *et al.*, 2006)), simultaneously infer a pathogen's spatial and evolutionary dynamics (phylogeography (Lemey *et al.*, 2009, 2010)) and infer population processes through time e.g. the effective population size (Ho & Shapiro, 2011).

Rabies virus is well suited to phylogenetic analyses, given a combination of its high evolutionary rate, lack of recombination and simple genomic structure. On a global scale dog RABV forms six major genetic *clades* (Fig. 2.2), five of which are associated with particular geographic regions: Africa 2, Africa 3, Arctic-related, Asian, and the Indian subcontinent (Bourhy *et al.*, 2008). Grouping of these clades largely reflects major barriers such as oceans, large mountain ranges and deserts, or historical colonisation events. The Arctic-related viruses are relatively well distributed, reflecting the lack of barriers in the far north, and a sixth clade has a cosmopolitan distribution, reflecting historical waves of human migrations and colonisations (Bourhy *et al.*, 2008; Smith *et al.*, 1992). Such efficient dissemination is likely to be a result of the virus' relatively long and variable incubation period occasionally facilitating long-distance transport of infected dogs.

The strong phylogeographic structure exhibited by most dog rabies clades is mirrored in wildlife rabies, where the same pattern is evident even at relatively small spatial scales (Biek *et al.*, 2007) and may be explained by a *"surfing mutation"* model (Excoffier & Ray, 2008). According to this model, new lineages arising during initial colonisation are able to reach high frequencies driven by an epidemic wave, whereas subsequent lineages do not benefit from such conditions and fail to infiltrate the dominating lineage clusters. Initial invasion events can therefore markedly influence the phylogeographic structure of rabies and, at least in wildlife, this structure can remain intact for decades (Biek *et al.*, 2007; Szanto *et al.*, 2011). However, with ongoing movement of infected individuals and the immigration of new lineages into an area, this structure would be expected to erode over time. How environmental variation affects these processes of emergence and subsequent erosion of phylogeographic structure are questions central to defining the landscape epidemiology of RABV and I will return to them throughout this review.

2.5 Landscape level effects on rabies dynamics

Since the transmission of rabies is entirely dependent on host movements initiating contact between infectious and susceptible individuals, the landscape that they occupy and disperse within heavily influences the ability of the virus to infect new hosts. Indeed, wildlife rabies epidemics tend to spread as irregular waves that differ in velocity according to heterogeneities in the landscape (Russell *et al.*, 2006). These "heterogeneities" include natural features such as mountain ranges, water bodies and deserts, but also anthropogenic features including roads and vaccine corridors. The influence of spatial heterogeneity on rabies spread can be broken down into three aspects, each of which are discussed in detail in the following sections: (1) host movements: natural versus human-mediated; (2) landscape attributes influencing rabies spread; and (3) population level effects.

2.5.1 The role of human vs. natural dispersal

Throughout history, humans moving animals has repeatedly led to the emergence and spread of rabies in susceptible populations, either through deliberate translocation (e.g. the relocation of animals, some of which may be incubating virus (Nettles *et al.*, 1979) or inadvertent movement (e.g. raccoons on garbage trucks (Wilson *et al.*, 1997).

Although human-mediated long-distance movement was found to be a significant feature of raccoon rabies spread (Smith *et al.*, 2002), it was still rare compared to natural dispersal, and stochastic in its occurrence. In contrast, human-mediated dispersal of dog rabies appears more widespread and potentially predictable due to possible links with human activities and migra-

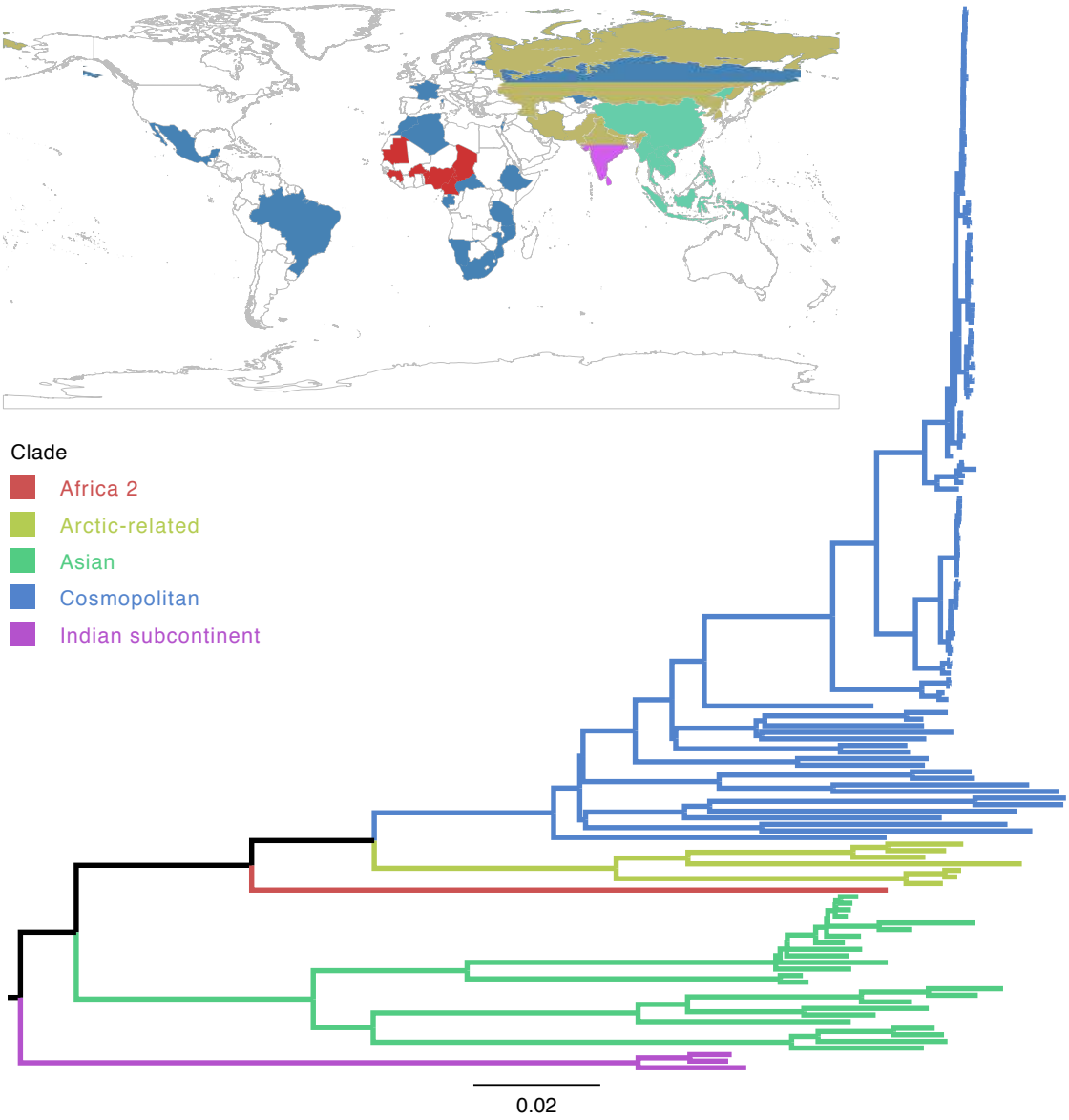


Figure 2.2: Global maximum likelihood phylogeny of rabies virus estimated from published whole genome sequences available in GenBank (listed in Appendix A) and additional sequenced Tanzanian RABV produced as part of this thesis (Appendix C) showing five of the six major clades (Africa 3 clade not shown as this is a mongoose-associated variant) and their global distribution. Global distributions were mapped according to the country of origin of the shown whole genome sequences and previously documented clade distributions described in (Bourhy *et al.*, 2008). Note both Arctic-related and Cosmopolitan clades have been found in Russia. Branch lengths are scaled by the number of substitutions per site.

tion patterns. The European colonisation of Africa is thought to have led to the introduction and subsequent expansion of canine rabies across the continent (Bourhy *et al.*, 2008; Lemey *et al.*, 2010; Talbi *et al.*, 2009); increasingly frequent records of re-introductions into countries where rabies has been eliminated have been recorded (Gautret *et al.*, 2011; Weiss *et al.*, 2009;

Table 2.1: Studies that have examined sources of spatial heterogeneity in dog rabies dynamics at a landscape scale; Gen= genetic data, Epi= epidemiological data. ML=Maximum Likelihood

Scale	**Sources of spatial heterogeneity**	**Data**	**Analytical methods**	**Key points/Summary**	**Reference**
Local (~5 yr period)	Individual host heterogeneity	Epi	Reconstruction of epidemic trees, outbreak simulations, construction of transmission networks based on a spatial infection kernel and generation intervals estimated from epidemiological data using ML.	Contact tracing data used to generate robust estimates of epidemiological parameters.	Hampson *et al.* (2009)
Local (~5 yr period)	Spatial configuration of populations	Epi	Patch-occupancy models.	Uses dog bite incidence records to explore metapopulation dynamics.	Beyer *et al.* (2011)
Regional (20+ yrs)	Socio-economic drivers	Gen	Bayesian inference of phylogeny, molecular clocks and demography.	Attributes phylogeographic patterns to economic growth and migration patterns of humans.	Carnieli *et al.* (2011)

continued ...

Scale	Sources of spatial heterogeneity	Data	Analytical methods	Key points/Summary	Reference
Regional (10+ yrs)	Translocation, road networks, wildlife hosts	Gen-Epi	Phylogenetic inferencePatterns suggest importance of wildlife hosts in this system.	Spread of rabies coincided with major highways, and indicated translocation events.	Coetzee & Nel (2007)
Regional (2 yrs)	Translocation	Gen	Antigenic characterization through monoclonal antibody profiling and molecular sequence comparison.	Identified the regional source of a rabid dog case through forensic epidemiological tracing.	David *et al.* (2004)
Regional (10 yrs)	Reservoir hosts, cultural drivers	Gen	Bayesian phylogenetics; parsimonious construction of transmission networks	Used statistical parsimony to construct most likely transmission networks. Also discovered the potential influence of social drivers on rabies phylogeographic structure (pastoralist vs. non-pastoralist community structure)	Lembo *et al.* (2007)

continued . . .

Scale	Sources of spatial heterogeneity	Data	Analytical methods	Key points/Summary	Reference
Local & Regional (10+ yrs)	Reservoir hosts, transmission clusters in wildlife	Gen-Epi	Bayesian phylogenetics ; generation of epidemic trees based on probabilities of links between possible progenitors and suspected cases weighted by spatio-temporal proximity, ML estimation of spatial infection kernel and generation interval distribution.	Most likely reservoir hosts (domestic dogs) inferred. Found evidence for short-lived chains of transmission in wildlife, compared to self-sustaining transmission in domestic dogs.	Lembo *et al.* (2008)
Regional (1 yr)	Road networks, population density	Epi	Spatial analysis of reported rabies using GIS. Directional spread of rabies based on mean centre of cases and a standard deviational ellipse weighted by date of cases.	Distribution of cases followed road network and towns with high human density and high numbers of free-roaming dogs.	Tenzin *et al.* (2010)

continued ...

Scale	Sources of spatial heterogeneity	Data	Analytical methods	Key points/Summary	Reference
Country (3 yrs)	Translocation, cross border incursions	Gen-Epi	Bayesian inference of phylogeny, molecular clock, demography and phylogeographic diffusion based on discrete spatial states.	Frequent incursions across country boundary; first molecular evidence for a long distance translocation of a rabies sub-lineage in Africa.	Hayman *et al.* (2011)
Between and within country (20+ yrs)	Geopolitical boundaries,translocation,road networks	Gen-Epi	Bayesian inference of phylogeny, molecular clock, demography and phylogeographic diffusion based on discrete spatial states; comparison of different geographic predictors of viral diffusion using model selection; quantification of phylogeographic clustering using association index; spatial simulation to test natural vs. human-mediated dispersal	Demonstrates and quantifies the anthropogenic influence on dog rabies dissemination. Best fit models implicate road networks as important predictors of rabies dispersal.	Talbi *et al.* (2010)

continued ...

Scale	**Sources of spatial heterogeneity**	**Data**	**Analytical methods**	**Key points/Summary**	**Reference**
Sub-continent (20+ yrs)	Geopolitical boundaries	Gen-Epi	Bayesian inference of phylogeny, molecular clock, demography and phylogeographic diffusion based on discrete spatial states.	Strong population subdivision at the country-level according to phylogeographic patterns.	Talbi *et al.* (2009)
Global (20+ yrs)	Mountain ranges, large water bodies, deserts, oceans	Gen-Epi	Bayesian inference of phylogeny, molecular clock, demography; and parsimony-based approach to determine the geographical structure of dog RABV phylogeny.	Elucidation of global patterns of dog rabies determined by major natural landscape barriers and historical colonisation events.	Bourhy *et al.* (2008)

Zanoni & Breitenmoser, 2003), and canine rabies has recently emerged on several previously rabies-free islands in Indonesia via fishermen importing incubating dogs (Susilawathi *et al.*, 2012; Windiyaningsih *et al.*, 2004). Even on smaller scales, phylogeographic patterns frequently imply translocation events that increase the level of mixing between lineages and may obscure patterns of lineage clustering (David *et al.*, 2004; Hayman *et al.*, 2011; Talbi *et al.*, 2010).

Although often inferred, it is difficult to quantify the influence of human-mediated movements on the diffusion of RABV. Though translocations may be frequent, the likelihood of establishment and spread is much lower (Smith *et al.*, 2005). But, unlike for wild animals, where translocation is risky and often unsuccessful, human-mediated displacement from a dog's original home range confers fewer risks. Dog translocations are therefore likely to be more successful in initiating new disease foci, and with potentially higher consequences, introducing disease into previously disease free areas. Given that long-distance movements instigated by humans are often in response to social drivers, the genetic structure of dog rabies is expected to reflect variation in cultures and socioeconomic factors. For example, rural workers in Thailand often relocate to urban areas where work is more readily available in the off-growing season, taking their dogs, and sometimes rabies, with them (Denduangboripant *et al.*, 2005). Similarly, the resurgence of canine rabies in KwaZulu-Natal in the 1970s was linked to refugee movements from Mozambique (Cleaveland, 1998; Swanepoel *et al.*, 1993) and phylogenetic patterns of RABV in parts of northern Tanzania can be tentatively explained by the movement of nomadic Maasai pastoralists with their dogs (Lembo *et al.*, 2007). Counterproductively, dog-owners have been reported to move their dogs to avoid the threat of culling during attempts to control rabies, which may explain more rapid spread of rabies than would be predicted by dog movement alone. Importantly, many of these aspects of human geography are quantifiable, making it possible in principle to generate testable hypotheses about their role in determining rabies phylogeographic structure.

Talbi *et al.* (2010) used spatial simulations to show that the observed patterns of spread in North Africa could only be explained by the occurrence of long-distance translocation events as opposed to natural dog movements alone. The wave-like patterns that are a signature of wildlife rabies and determined by natural host movements are less evident in domestic dog rabies, and it is unknown to what extent natural dog movements vs. human mediated movements determine the observed phylogeographic patterns. A recent study in a Kenyan rangeland found that healthy domestic dogs rarely moved more than 50 metres from their home bases, and the maximum distance recorded was 3.2 km (Woodroffe & Donnelly, 2011) & see Fig. 2.3). In contrast, dispersal distances in wildlife hosts tend to be at least an order of magnitude higher (Cullingham *et al.*, 2008). However, rabies infection leads to behavioural changes, which may alter movement patterns; in northern Tanzania, most contacts with rabid dogs occurred within a kilometre of an animal's homestead (mean ˜0.88 km), but a small proportion of rabid dogs ran over 15 km while infectious (Hampson *et al.*, 2009). Dog

movements also vary depending on the societal context and livelihoods of the communities to which they belong (Woodroffe & Donnelly, 2011). For their dispersal simulations, Talbi *et al.* (2010) used measures of natural dog movements based on a specific locality in Tanzania (Hampson *et al.*, 2009) but it remains unclear whether such data are transferrable among different geographic localities. Research on domestic dog ecology and movement is lacking and detailed, location-specific data are required in order to reliably test hypotheses for alternative drivers of viral diffusion.

2.5.2 Landscape attributes influencing rabies spread

As a parasite transmitted through direct host-to-host contact, spatially-defined genetic discontinuities in RABV populations may indicate a barrier to host contact (Biek & Real, 2010). I consider a "dispersal barrier" to be any spatial feature of the landscape that impedes the gene flow of a pathogen. Spatially explicit models incorporating landscape features can help uncover barriers to gene flow based on slower progression than expected over a homogeneous area. However, the interpretation of genetic discontinuities requires caution as phylogeographic patterns may arise as a result of historical colonisation events and irrespective of physical barriers (Real & Biek, 2007; Talbi *et al.*, 2009).

Smith *et al.* (2002) demonstrated a seven-fold reduction in rates of raccoon rabies spread in North America due to the presence of rivers and forest cover. Similarly, the Vistula River separated distinct clusters of a red fox variant and raccoon dog/fox variant in Europe (Bourhy *et al.*, 1999). Clearly, large water bodies impede host movement, but have differential success as barriers according to additional factors such as habitat suitability Smith2005,Cullingham2009, physical geography (Rees *et al.*, 2009), and width and flow rates (Bourhy *et al.*, 1999). In an attempt to assess the differential permeability of barriers, Rees *et al.* (2008) used genetic simulation modelling to assess the effect of the Niagara River on rabies spread from New York State to Ontario. Comparing genetic population structure derived from field data, with simulated population expansion scenarios they ascertained a 50% barrier effect. In contrast, homogeneous landscapes that lack environmental barriers are particularly vulnerable to the rapid expansion of an introduced pathogen, as demonstrated by the increased speed with which raccoon rabies is predicted to cross central Ohio, which lacks major natural barriers, compared to neighbouring states (Russell *et al.*, 2005). With the exception of major landscape features inferred as barriers to dog rabies on a global scale (Bourhy *et al.*, 2008), we know relatively little about barrier effects at smaller scales. Determining whether natural barriers prevent the dissemination of dog RABV or if ties to human ecology negate their effect is an important research question with significant implications for control.

Political borders are another potential form of barrier that may play a role in the containment of dog rabies, but have no evident impact on wildlife rabies dispersal. For example, distinct

Box 2.3: Phylogenetic inference: Bayesian and maximum likelihood estimation.

Estimating phylogenetic trees has become a standard means of analysing sequence data. The availability of explicit models of molecular evolution enables tree reconstructions that explore a range of mutational pathways between sequences and statistical measures can be used to estimate the "best" tree and measure uncertainty in the assignment of branches and subtrees. There are many tree-building methods that can be implemented, many of which involve using a frequentist or Bayesian approach to infer and ascertain the quality of the phylogeny.(Huelsenbeck & Rannala, 2004). While these approaches are based on the use of the same underlying evolutionary models they undertake different approaches to assess phylogenetic uncertainty. See Figure I for an overview of both approaches described below.

a) Maximum likelihood is an example of a frequentist approach, which involves evaluating all possible mutational pathways under an explicit model of sequence evolution to estimate trees and their probability of generating the observed sequence data (likelihood). The summary of this data is the tree with maximum likelihood. Confidence is assessed by bootstrapping, which works by randomly resampling characters in columns of the sequence alignment with replacement, rebuilding the tree and measuring the frequency that the same phylogenetic groupings are recovered. Resampling is usually undertaken 100 or 1000 times (Burr *et al.*, 2002) with a value $\geq$70% (i.e. the grouping was recovered in 70% of the resampled data) generally accepted as dependable support for a group (Hillis & Bull, 1993).

b) Bayesian inference methods are less computationally intensive, taking advantage of Markov chain Monte Carlo (MCMC) algorithms to evaluate a reliable sample of trees(Huelsenbeck *et al.*, 2001). Inference is based on the calculation of a posterior probability, which is a quantity reflecting the confidence that that the tree is correct, assuming that the model is correct (Huelsenbeck & Rannala, 2004). This is estimated given a prior probability of a tree (usually all trees are considered equally probable *a priori*) which is updated by combining with the likelihood of observed data using Bayes' Theorem (Huelsenbeck *et al.*, 2001). As solving this analytically for all possible trees is too complex, Markov chain Monte Carlo algorithms are relied upon to approximate posterior distributions. Confidence in the tree or subtree branching is assessed by examining the posterior support, with 90% indicating strong support.

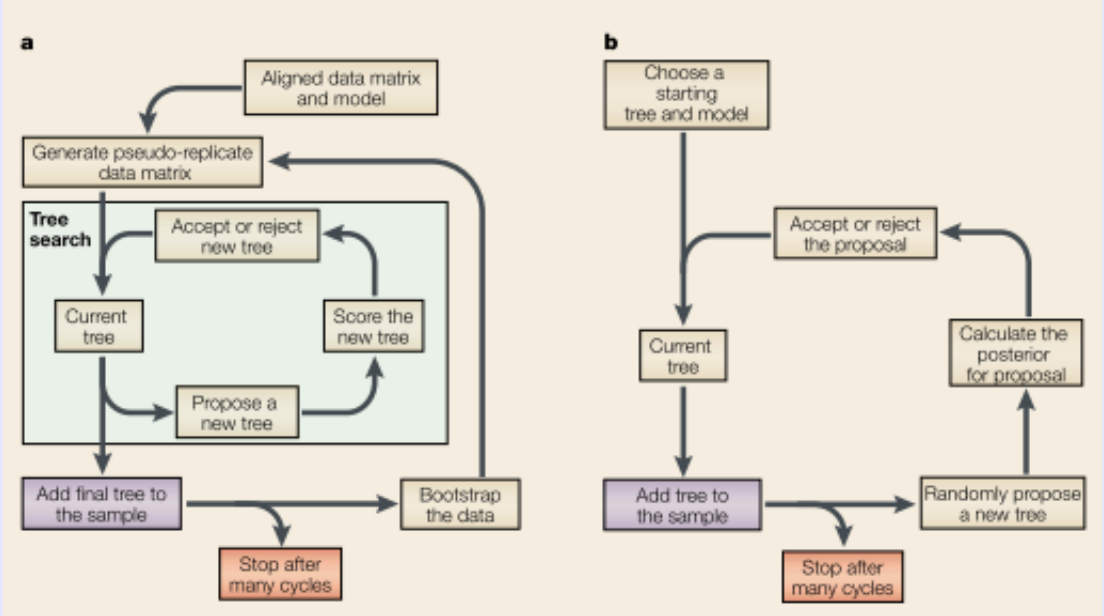


Figure I: Maximum likelihood (a) and Bayesian (b) approaches to infer a phylogeny and ascertain phylogenetic confidence. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics. (Holder & Lewis, 2003), copyright 2003

and almost monophyletic RABV groups associated with North African countries indicate restricted movement across geopolitical boundaries (Talbi *et al.*, 2010). This well defined population structure at a regional scale contrasts with a relatively fluid dissemination within countries. On the one hand, this suggests that country level vaccination programmes should have a good chance of eliminating dog rabies even in contiguous landscapes. However, these findings are not generalisable, with epidemiological analyses of RABV in Ghana highlighting frequent cross-border incursions (Hayman *et al.*, 2011). In addition, time-series analyses indicate large-scale synchronous dynamics of rabies across multiple countries in eastern and southern Africa, possibly due to a combination of human or wildlife-mediated long-distance dispersal and a lack of sustained control programmes (Hampson *et al.*, 2007). Understanding the circumstances whereby political boundaries act as dispersal barriers should provide guidance for whether control programmes require regional co-operation or can be sustained at a national level with appropriate border controls. As a first step, it would be useful to compare phylogeographic structure of dog RABV among different parts of the world and across hierarchical spatial scales. Objective measures for such a comparison can be obtained by quantifying the degree to which sequences cluster on a phylogeny according to their geographic location. Several statistics are available for this (Parker *et al.*, 2008), of which the *association index (AI)* (Wang *et al.*, 2001) has been found to be of particular utility. This measure of phylogeny-trait correlation may provide an initial descriptive analysis of consistency or variability in the hierarchical genetic structure of dog rabies between continents or countries that could determine the generalisation of results across different systems. For example, we might hypothesise that regions with large population densities and a high level of transport infrastructure, e.g. parts of Asia, show less distinctive phylogeographic patterns (a low AI value) compared to more sparsely populated and less developed landscapes like those in sub-Saharan Africa (high AI value).

Quantifying barrier effects, to predict the likelihood of incursions or to exploit them for control programmes, is an important area yet to be tackled for dog rabies. The impact of anthropogenic landscape features on spread, at least in a qualitative sense, has been noted in several countries. For example, Tenzin *et al.* (2010) mapped the spread of rabies in Bhutan showing a strong visual pattern alongside road networks and towns with high dog-to-human ratios; and phylogeographic patterns in north Mexico and Thailand match the distribution of major migration routes (De Mattos *et al.*, 1999; Denduangboripant *et al.*, 2005). It can be hypothesised that features important to wildlife rabies become less significant to dog rabies as anthropogenic features take over the landscape and mediate the effect of barriers, i.e. bridges, roads and transportation make it possible for dogs to circumvent natural barriers and support a relatively fluid dissemination of rabies across naturally heterogeneous landscapes. These hypotheses remain largely untested and thus present a fruitful area for landscape genetics investigations.

Uncovering those features that facilitate viral spread is an equally important aspect of assess-

Figure 2.3: Varying spatial complexity in areas with endemic dog rabies as a result of increasing dog population density, A) low density: Ngorongoro District, Tanzania; B) medium density: Serengeti District, Tanzania, C) high density: Hermosillo,Mexico. Red circles highlight settlements in rural areas.Maps obtained using Google Earth (`http://earth.google.com`).

ing the landscape. As previously discussed, anthropogenic effects may facilitate the transmission of rabies across a larger scale than natural movements alone allow. The impact of human-mediated dog transport may by explored by phylogenetic analysis, with translocations implicated by the presence of a cluster-specific variant in a distant locality (see David *et al.* (2004) and Cohen *et al.* (2007), Table 2.1), essentially an assignment approach (Paetkau *et al.*, 1995). While this method is not a definitive measure of a translocation, it can identify the most likely dispersal scenario. Talbi *et al.* (2010) used Bayesian phylogeographic diffusion models (Lemey *et al.*, 2009) in an attempt to quantify the importance of various anthropogenic predictors on the observed spread of rabies cases in North Africa (see Table 2.1). Model fitting indicated that dispersal patterns among towns were best explained by road distance, consistent with the anticipated role of human movement. Interestingly, further refinements to calculating distance matrices, such as an accessibility index based on road type and travel time, received only limited model support and road distances were only a marginally better predictor than great-circle distances. This may indicate that RABV dispersal follows a rather homogeneous spatial diffusion process, without any particular effect of landscape heterogeneity. More likely however, it means that more useful geographic predictors of human- mediated dispersal have yet to be found. While this was not possible in the Talbi *et al.* (2010) study, their method can accommodate a wide range of geographical and environmental predictors and thus provides a promising general framework for examining landscape effects in future data sets.

2.5.3 Population level effects and metapopulation dynamics

Infectious disease dynamics are often described in terms of a *metapopulation*, with host populations divided into smaller, spatially structured sub-populations that exist with different inter- and intra-patch dynamics (Grenfell & Harwood, 1997). Heterogeneities in the spatial configuration of host populations are critical to understanding the persistence of endemic pathogens (Hagenaars *et al.*, 2004) and incorporating this social/spatial structure is particularly important for developing mechanistic models of acute-acting infections like rabies (Cross *et al.*, 2005). In the case of dog rabies, human settlements (village/town/city etc.) can be considered habitat patches that vary in host density and connectivity (Fig 2.3). For example, rural areas typically exist as an array of villages connected to larger towns and cities by major roads. Fragmentation and low connectivity, i.e. long distances between settlements, natural barriers and lack of transportation networks, likely restrict the ability of a dog to move across a landscape and, hence, limit rabies spread. Hypothetically, limited contact between dog populations and correspondingly strong spatial structure of rabies would be expected in the least developed regions where travel is most limited.

Local within-patch dynamics are potentially a key component contributing to the persistence and maintenance of disease at larger scales. Therefore, it is important to quantify the relative

contribution of different patches to overall disease persistence in a system. This might involve characterising patches based on dog density, turnover and growth rates, levels of ownership and vaccination, alongside measures of connectivity. For example, dog turnover rates can be extremely high e.g. the Machakos region in Kenya has a dog population estimated to grow by 9% per annum, making rabies control difficult in this area (Kitala *et al.*, 2001). Areas such as these may possibly be identified as hotspots or source patches where rabies has a high chance of being maintained and spreading from.

A potentially useful tool for quantifying how rabies dynamics differ among local patches is the basic reproductive number, R_0. R_0 quantifies transmission potential at the beginning of an epidemic and is defined as the average number of secondary cases resulting from a single infectious individual in a completely susceptible population. In theory, $R_0 > 1$ is required for successful invasion and spread. In an endemic context or to measure R through time a similar metric known as the effective reproductive number, R_t, can be used to measure the number of secondary infections resulting from a single infectious individual at time t (Anderson & May, 1991). This can be useful to quantify time-dependent transmission potential through the course of an epidemic or the impact of control interventions i.e. R in a partially immune population. Similarly, R_t must be held above 1 in order for the disease to spread and persist.

Theory would suggest that for directly transmitted diseases such as rabies, R_0 should increase with host density. However, empirical evidence for density-dependent dynamics is equivocal, with no detectable differences in R_0 among dog populations with varying densities around the world (Hampson *et al.*, 2009). While it is unclear why R_0 is so insensitive to differences in dog density, e.g. Mexico (~100 dogs/km^2; Eng *et al.* (1993)) versus rural Africa (~10 dogs/km^2; Lembo *et al.* (2008)), this finding is more consistent with a frequency-dependent mode of rabies transmission. For both density and frequency-dependent disease dynamics stochastic extinctions are expected once density drops below a certain level, meaning that low-density patches may act as barriers to spread. Indeed, rabies has been observed to spread but appears less able to persist in low density dog populations, e.g. Ngorongoro in Tanzania (average of 4.2 dogs/ km^2; Lembo *et al.* (2008)), suggesting that neighbouring source populations are required for rabies maintenance. Cross *et al.* (2007) discuss the incorporation of heterogeneities in population structure by expanding on the utility of R_0 as a measure of disease transmission. R_0 assumes that the population is evenly mixed, but the hierarchical nature of disease invasion requires stochastic models that incorporate both within-patch transmission (R_0) and the factors contributing to persistence, i.e. between patch transmission- namely the recruitment of susceptibles, group size and the infectious period. These questions of potential metapopulation structure, patch variability and source-sink dynamics (Pulliam, 1988) represent a particularly pertinent and rich area for future studies of RABV which has yet to be explored using genetic approaches and are vital to inform efficient control strategies.

Host movement rates and connectivity between sub-populations are crucial predictors of dis-

ease invasion (Beyer *et al.*, 2011; Cross *et al.*, 2005). Epidemiological data available at a localised scale, e. g. the incidence of dog bites, can provide a simple spatio-temporal measure for metapopulation analysis, as shown by Beyer *et al.* (2011). Patch-occupancy models identified two metrics of connectivity, the distance between neighbouring villages and the size of villages receiving infection, as significant factors facilitating the transmission of disease. From a population genetics perspective, coalescent analysis of sequence data could provide the means to identify metapopulation dynamics that may not be discernible using epidemiological data alone. Such techniques have been successfully applied, for example, to understand the persistence of influenza A virus, with molecular clock-based estimates of divergence times of *most recent common ancestors (MRCA)* and demonstrating persistence due to dynamic migration patterns rather than source-sink dynamics (Bedford *et al.*, 2010; Russell *et al.*, 2008).

More generally, the fine-grained spatial genetic structure evident in wildlife rabies that is lacking for dog rabies may be due to differences in time frame. The best-studied wildlife rabies dynamics come from invasions that originated a few decades ago (Biek *et al.*, 2007; Bourhy *et al.*, 1999; Real *et al.*, 2005a), compared to endemic foci in domestic dogs, thought to have persisted for centuries (Bourhy *et al.*, 2008). Factors influencing persistent endemic cycles are likely to differ from those determining epidemic expansions, with viral population turnover, increased mixture of lineages and the greater role of human-mediated movement resulting in less clearly defined fine-grain viral structure. Again, metapopulation models that capture population connectivity at appropriate spatial and temporal scales may shed light on the genetic structure of endemic foci. Characterising the landscape connectivity between rabies endemic areas (or individual rabies cases, given the resolution of spatial data) may be facilitated using techniques applied in conservation genetics that have yet to be exploited for disease dynamics. For example, the program Fractionnator, (`http://www.unil.ch/biomapper/frictionnator/frictionnator.html`), provides a "strip statistic" quantifying the effect of landscape features on the relatedness of all possible pairs of individuals sampled. This entails defining grid cells within a strip across the sampled landscape and assessing the abundance of landscape features within each strip. In addition, advances in the development of spatial analysis software such as python-based customised GIS tools (e.g. Etherington, 2011) that allow the visualisation and measurement of genetic relatedness and landscape connectivity based on least cost path analysis, R packages such as Maptools (Lewin-Koh *et al.*, 2012) and alternative open-source software such as PASSaGE 2 (Rosenberg & Anderson, 2011) provide utilities for measuring connectivity between patches, quantifying patterns in spatial data and potentially identifying barriers to dispersal.

Elucidating how heterogeneity at the host level contributes to dynamics at the patch level is an important aspect of understanding the mechanisms underlying rabies persistence and spread. Differential within-patch dynamics resulting from individual heterogeneity including variation in biting frequency, host population size, the structure of a settlement, may confer

properties that promote the effectiveness of a patch as a source of infection. Tracing transmission pathways may provide a further means of elucidating source populations that initiate chains of infection, particularly at very fine-scale resolution. Lembo *et al.* (see 2007, Table 2.1) used parsimonious transmission networks to infer reservoir host dynamics and patterns of interspecific transmission, but this technique could also be used more generally to uncover dispersal patterns, e.g. if a particular patch repeatedly acts as a source of infection. The construction of transmission trees from epidemiological data on the timing and locality of rabies cases is another useful approach for identifying transmission pathways and can be supplemented by contact tracing data (Lembo *et al.*, 2008). Contact tracing of rabies is aided by the memorable nature of rabies bites and therefore presents a unique situation where the collection of contact data is achievable. However, building transmission trees in this way becomes increasingly unreliable as larger numbers of missing links are inferred, and these models may also simplistically predict unlikely transmission events if knowledge of the landscape is not incorporated into algorithms e.g. predictions of transmission across major landscape barriers. Genetic data could be used to quantify mutation rates and build independent time-scaled transmission networks against which epidemiologically constructed transmission trees could be calibrated. Ultimately, transmission trees will be most effectively constructed with the combined use of genetic and epidemiological data, and appropriate datasets are increasingly becoming available for rabies, but further computational and statistical advances are required.

Host transmission is typically heterogeneous, often described by the 80/20 paradigm in which individuals differentially carry and transmit pathogens i.e. 80% of infections are carried by 20% of the population (Woolhouse *et al.*, 1997). In the most extreme case this heterogeneity can exist in the form of *superspreaders* (Lloyd-Smith *et al.*, 2005). There have been no in-depth studies of superspreading in domestic dogs carrying rabies and the contribution of superspreaders to rabies spread and persistence is controversial (e.g. Hampson *et al.*, 2007) vs. (Talbi *et al.*, 2009), but large variations in biting frequency have been observed (Hampson *et al.*, 2009). An important distinction must be made between superspreading in a restricted area, e.g. the same village, and a "spatial superspreader" that has dispersed infection over large distances and multiple settlements (see Fig. 2.4 for visual example). Although the number of secondary cases caused by such individuals may be equivalent, spatial superspreaders are more important from an epidemiological perspective if they are responsible for connecting sub-populations. Tracing resulting infections to their source through genetic sequence data may identify hotspots where certain patch features promote greater dispersal ability i.e. spatial superspreading events (as suggested by (Cross *et al.*, 2007)). Ideally, this would utilise retrospective genetic tracing of networks of infection, similar to that used for identifying source infections in foot-and-mouth disease (FMD) (Cottam *et al.*, 2008). Statistically robust methods are yet to be conceived, but accounting for missing links in transmission chains is an important area for future development. As a major advantage compared to FMD, contact that may have lead to transmission is much easier to define for rabies. Contact tracing data, if available, could therefore complement genetic data to offer a more comprehensive view of

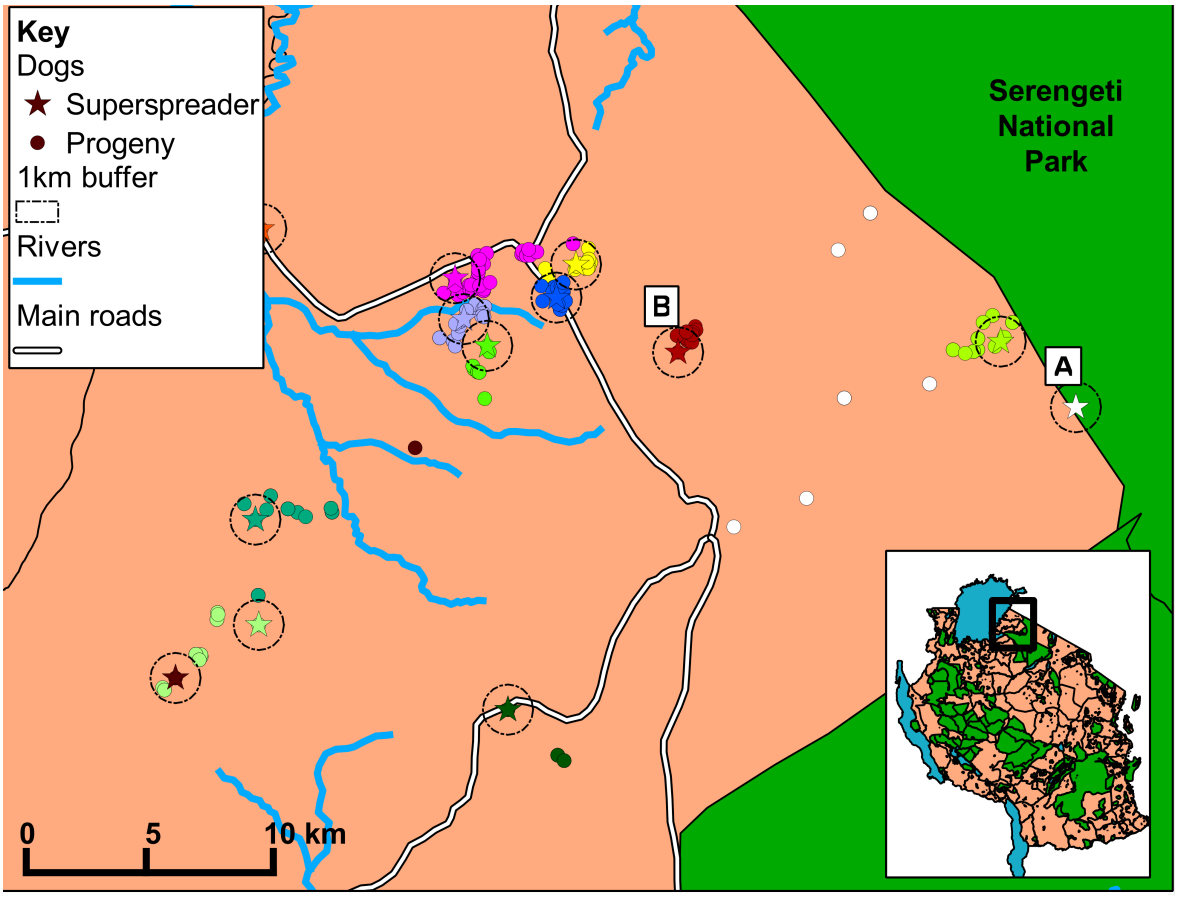


Figure 2.4: Dispersal of bites from superspreading dogs resulting in rabies transmission in an area of the Serengeti District in Tanzania. Roads and rivers are shown to highlight the potential influence of landscape features on the dispersal of rabies- tentative observations indicate that superspreader progeny appear to cluster alongside roads and movement may be restricted by the presence of rivers (but other landscape features not shown may also be responsible for influencing dispersal patterns). Two potential types of superspreader are also highlighted in the map: A) a spatial superspreader, which transmits over a large spatial area, potentially connecting sub-populations and may be important from an epidemiological perspective; and B) a superspreader with a limited dispersal range that infects a large number of progeny but remains within a small spatial radius. Inset map shows the location of the Serengeti District within Tanzania. prevention

how individual heterogeneity can impact on dynamics.

2.6 Integrating landscape epidemiology into rabies control

Landscape genetics has undoubtedly generated fundamental insights into the dynamics of rabies but has arguably yet to make major contributions to its control. The preceding discussion illustrated potential avenues for exploration and here I aim to more clearly define a landscape genetics research agenda that could directly benefit the planning or implementation

of programmes that aim to control or eliminate rabies.

Considerable progress has been made in the development of blueprints and operational toolkits for rabies control (see Blueprint for rabies prevention and control, August 2010, `http://www.rabiesblueprint.com/`, and Lembo *et al.* (2011). The rabies blueprint presents a major step in the global fight against rabies, providing guidelines and economically feasible strategies to aid policy makers and local communities seeking to embark on rabies intervention and control measures. Landscape genetics research constitutes an important resource within this multidisciplinary approach to help advise and sustain successful control initiatives.

Vaccination is widely deemed the most effective means of rabies control with demonstrated successes in reducing incidence and eliminating disease even in areas with limited resources (Cleaveland *et al.*, 2003; Lembo *et al.*, 2010; Schneider & Leanes, 2007). Molecular genetics has demonstrated that domestic dogs are critical reservoirs for canine rabies, even in parts of Africa with abundant wildlife populations, and therefore indicates that controlling dog rabies through vaccination should eliminate "spill-over" infections in humans and other animal populations throughout Asia and Africa (Lembo *et al.*, 2008, 2010, 2007). By changing the susceptibility of populations, mass vaccinations change the landscape in which rabies circulates, and at sufficiently high levels of vaccination coverage, transmission can be interrupted. There are two mechanisms by which vaccination alters the landscape to control rabies. The first is that vaccination itself creates a barrier of susceptible individuals that block the dispersal of the pathogen across the landscape, and the second is via the reinforcement of existing barriers with vaccination. Using landscape genetics to explore the potential impacts of these aspects of landscape control is a logical next step.

Recurrent rabies epidemics occur across large areas where vaccination programmes are patchy and unsustained (Hampson *et al.*, 2007), suggesting that a proactive long-term vaccination programme coordinated across political boundaries is required for success. Indeed, the effectiveness of such a programme has been proven by the intensive vaccinations coordinated by the Pan American Health Organisation (PAHO) in Latin America over the past few decades: dog rabies has been eliminated in a large portion of the southern continent, and reported cases from other countries are highly localised due to restrictions on trans- mission pathways from vaccination barriers (Schneider & Leanes, 2007). Crucially, cases of human rabies have dropped in these areas, reinforcing the importance of effective dog rabies control strategies. Strategically placing a vaccine barrier could maintain freedom from rabies resulting from successful interventions in otherwise landlocked areas.

Environmental or anthropogenic barriers to natural transmission of rabies offer the opportunity to strengthen and smarten vaccination initiatives. Vaccination can be viewed as a form of barrier that impedes rabies spread, and therefore many of the techniques used to draw insights on the permeability of barriers could equally be applied to vaccination programmes. *Oral Rabies Vaccination (ORV)* campaigns for wildlife, including bait distribution in prox-

imity to a pre-existing natural barrier (the Appalachian Mountains), and utilising existing environmental features to reinforce control campaigns (Wandeler *et al.*, 1988) have successfully contained wildlife rabies. Past experience has shown vulnerability to breaches associated with the differential permeability of barriers and emphasises the need for ongoing and targeted surveillance to enable early detection and swift responses to incursions (Russell *et al.*, 2005). Yet despite this, few studies have quantified the utility of barriers within a landscape, and there are no guidelines available for the design and implementation of effective cordon sanitaires for dog rabies.

In addition to barrier studies, modern application of spatial data is allowing us to make increasingly accurate measurements of epidemiological parameters that may affect disease dynamics. For example, Bharti *et al.* (2011) demonstrate the use of remote sensing to test for human predictors of disease. They used anthropogenic light from satellite imagery as a measure of seasonal fluctuations of human populations. The observed fluctuation in light intensity (as a measure of population density) correlated to measles distribution and spread in cities in Niger, and provided an accurate, near real-time representation of short-term population fluctuations that may drive pathogen transmission. This approach demonstrates the use of relatively simple proxies for quantifying migration patterns in poorly resourced regions, often the same regions carrying the highest disease burden from dog rabies.

Large-scale interventions are expensive, and adaptive management is often required alongside intervention. Refinements in resource management will ultimately rely on a combination of knowledge from genetics, landscape and host ecology as part of a reactive programme. Specific landscape elements affecting dog rabies spread are generally not known *a priori*, so exploration of phylogeographic patterns may identify genetically distinct viral or host populations in areas, which could be targeted for vaccination. However, experience from wildlife rabies suggests that caution is warranted when taking such an approach. Firstly, apparent boundaries between areas dominated by different genetic lineages may have emerged during the initial invasion process and thus do not necessarily reflect areas of low permeability for the virus (Biek & Real, 2010; Real *et al.*, 2005b). Secondly, while the stability of these phylogeographic domains indicates a lack of mixing between them, this simply suggests that immigrating viruses find it difficult to invade areas with an already established focus, but does not signify the absence of viral immigration per se. Such areas with putatively self-contained endemic foci could therefore experience a high risk of rabies re-emergence following successful eradication, unless vaccination effort remains high. Whether the second consideration equally applies to dog rabies is currently not clear as pertinent empirical studies are lacking. As with many research problems in landscape genetics, these types of question may be very productively studied using simulation tools (Epperson *et al.*, 2010). Simulations may indicate the most effective location for vaccine corridors/barriers e.g. vaccination on the far side of a natural barrier (Russell *et al.*, 2006), whereas metapopulation models may elucidate areas with high connectivity that could be the source of persistence in endemic areas. Quantifying the degree of connectivity between

sub-populations is of critical importance to understanding how rabies is maintained across landscapes, and is an integral part of designing effective control interventions and cordons sanitaires that limit and contain the virus.

Ongoing surveillance is crucial to the long-term success of rabies control, and considerable value can be added to surveillance initiatives through the incorporation of landscape genetics, with the potential to determine the source of incursions and reveal transmission pathways. Epidemiological surveillance may facilitate active case detection and identification of circulating strains, helping to identify areas missed by vaccination. Using a molecular genetics approach during the 2007 FMD outbreak in the UK, allowed swift and effective containment of the outbreak by directing interventions to the critical areas (Cottam *et al.*, 2008). For countries with limited resources, surveillance in areas with ongoing control programmes and retrospective analysis may be useful for identifying remaining foci of infection to be targeted by vaccination, sources of incursions, or spillover events into or from wildlife. In addition, metapopulation models may be utilised to predict the direction of spread of rabies in naïve populations (Beyer *et al.*, 2011), directing the location of sentinel points and control measures to "hotspots" of infection (Haydon *et al.*, 2006) and putative transmission networks based on sampled cases should indicate how well current levels of surveillance are capturing rabies incidence based on inferred missing links between sampled cases.

Rabies containment and control will require ongoing surveillance and sustained control efforts. In time, when rabies incidence is reduced to low levels (the period when control measures often lapse), genetics can provide a means of directing resources to where they are most needed, maintaining a cost-effective approach. Given the small genome of RABV (12Kb), and advances in NGS sequencing, future exploration of phylogeographic patterns utilising whole genome sequencing is a realistic prospect. This promises to provide appropriately fine genetic resolution for samples collected on small spatio-temporal scales that might otherwise be uninformative. Application of novel techniques in landscape genetics to other RNA viruses such as FMD and Influenza highlights interesting possibilities for uncovering disease dynamics and aiding control directives (Bedford *et al.*, 2010; Cottam *et al.*, 2008). The technology and analytical power are available but have yet to be fully exploited for dog rabies, and thus offer exciting prospects on what can be achieved with their implementation in the future.

2.7 Conclusions

There is evidently a wide scope for the use of landscape genetics to explore and understand the dynamics of pathogen spread and persistence. Dogs are the principal reservoir of rabies, responsible for the majority of human rabies cases, yet we know little about the dynamics of the pathogen in this host. In order to reduce the many thousands of human rabies deaths that occur annually due to contact with rabid dogs, it is crucial that we uncover the mechanisms

governing viral dispersal across the landscape so as to direct successful control interventions. The foundation of knowledge from studies in wildlife populations provides a starting point but aspects of host behaviour, the inherent influence of humans, and the long-term endemic nature of dog rabies foci requires a modified approach be taken for dog rabies. There lies great potential for advancing the effectiveness of control campaigns in areas burdened with disease; specifically landscape genetics has most to contribute to improving surveillance and modifying control strategies based on information gleaned from surveillance. This includes the design and placement of cordons sanitaires, the prioritisation of effort towards persistent foci or targeting sources of outbreaks and conduits of transmission. Given the availability of powerful new genetic and spatial techniques, efforts now need to push towards real-world application of landscape genetics to rabies control and elimination.

CHAPTER 3

Elucidating the phylodynamics of endemic rabies virus in eastern Africa using whole-genome sequencing

3.1 Abstract

Many of the pathogens perceived to pose the greatest risk to humans are viral zoonoses, responsible for a range of emerging and endemic infectious diseases. Phylogeography is a useful tool to understand the processes that give rise to spatial patterns and drive viral dynamics. Increasingly, whole genome information is being used to uncover these patterns but it is unclear how much resolution can be achieved and down to what scale. Here, whole genome resolution was used to uncover fine-scale population structure in endemic canine rabies virus circulating in Tanzania, providing information that could be used to guide interventions, such as the spatial scale and design of dog vaccination campaigns and dog movement controls. This is the first whole genome population study of rabies virus and the first comprehensive phylogenetic analysis of rabies virus in East Africa, providing important insights into rabies transmission in an endemic system. In addition, sub-continental scale patterns of population structure were identified using partial gene data and used to determine population structure at larger spatial scales in Africa. While rabies virus has a defined spatial structure at large scales, increasingly frequent levels of admixture were observed at regional and local levels. Discrete phylogeographic analysis revealed long-distance dispersal within Tanzania, which could be attributed to human-mediated movement and I found evidence of multiple persistent, co-circulating lineages at a very local scale in a BEAST district, despite on-going mass dog vaccination campaigns. This may reflect a more continuous dispersal dynamic in endemic landscapes where rabies virus has been circulating for decades alongside increased admixture due to human-mediated introductions. These data indicate that successful rabies control in Tanzania could be established at a national level, since most dispersal appears to be restricted within the confines of country borders but some coordination with neighbouring countries may be required to limit transboundary movements. Evidence of a more dynamic diffusion process within Tanzania necessitates the use of whole genome sequencing to uncover finer scale population structure that can inform more targeted interventions to achieve and maintain freedom from disease.

3.2 Introduction

The general trend of increasing incidence and expansion of emerging or re-emerging zoonotic diseases (e.g. Ebola, Chikungunya, avian influenza) Greger (2007); Jones *et al.* (2008); Woolhouse (2002) and persistence of established zoonoses, such as canine rabies, highlights the ongoing challenges faced as we attempt to characterise and control them. The processes that drive the spread and persistence of infectious diseases are reflected in a genetic signature in pathogen genomes Biek & Real (2010). Understanding the processes that give rise to spatial population structure in pathogens can inform the management and control of infectious diseases. For example, analyses of evolutionary, epidemiological and ecological data have re-

cently demonstrated that global live swine trade strongly predicts the global dissemination of influenza A viruses in swine Nelson *et al.* (2015) and air travel has been revealed as a major factor driving the intra-continental spread of Dengue virus Nunes *et al.* (2014). Viral pathogens, particularly fast-evolving RNA viruses, are model systems to explore pathogen populations as they rapidly accumulate genetic diversity on a timescale similar to epidemiological processes Biek *et al.* (2015); Drummond *et al.* (2003). Statistical phylogeographic approaches are available Bedford *et al.* (2014); Bielejec *et al.* (2014); Lemey *et al.* (2009) to develop a quantitative understanding of the processes that give rise to spatial patterns in RNA viruses Holmes & Grenfell (2009) on epidemiological time scales. Whole genome sequencing (WGS) is increasingly being used as a means to extract these patterns but it is unclear how much resolution can be gained and at what temporal and spatial scale. In this chapter I use canine rabies virus (RABV) as a model to determine the spatio-temporal patterns of an endemic zoonotic virus using whole genome data to distinguish structure at an increasingly fine scale.

Box 3.1: Glossary

Bayes Factor (BF): the ratio of the marginal likelihoods of two models,used as a means of Bayesian model comparison (Drummond & Rambaut, 2007). A BF of 3 is considered the minimum value indicating positive support for a model (Kass & Raftery, 1995).

Bayesian stochastic search variable selection (BSSVS): process that estimates the posterior probability that an explanatory variable should be included in a model (Bloomquist *et al.*, 2010). In a phylogeographic model many of the transitions in the transition rate matrix used to model diffusion are unlikely to occur such that *a priori* many are suspected to be zero. BSSVS determines which rates are zero by testing variables in a linear regression model via an indicator function to select a parsimonious parametrisation of the rate matrix, uncovering the most likely migration patterns according to evidence in the data (Lemey *et al.*, 2009).

Bayesian skyline: a non-parametric coalescent method to estimate changes in past population dynamics by estimating a product of the effective population size and the time between generations, N_{et}. Skyline processes split the timeline into a number of intervals where the effective population size is constant within but variable between intervals(Drummond *et al.*, 2005).

Brownian diffusion (BD): in a phylogeography context diffusion can be modelled as a Brownian random walk, which is a stochastic process on a geographic surface in with stationary, independent increments that are normally distributed with mean zero and variance that scales linearly with duration (Faria *et al.*, 2011).

Continuous time Markov chain (CTMC): a stochastic process that emits discrete outcomes as a continuous function of time (Lemey *et al.*, 2009). The CTMC is a memoryless process that makes state transitions dependent only on the present state and independent of past behaviour, with waiting times between transitions determined by an exponential distribution (Bloomquist *et al.*, 2010). The CTMC is characterised by an infinitesimal rate matrix (Lemey *et al.*, 2009).

Effective population size N_e: the size of an idealised population (without selection or population structure) that experiences the same level of genetic drift as the studied population. N_e is usually lower than the actual population size N.

Relaxed random walk (RRW): relaxed version of strict Brownian diffusion where rate heterogeneity across the phylogeny is enabled by rescaling the variance along each branch using a scalar drawn independently from a specified distribution (Lemey *et al.*, 2010).

Robust counting: framework to count labelled state transitions (Markov jumps) in discrete evolutionary traits over the course of CTMC path, while protecting against bias due to model misspecification of the underlying CTMC (Brien *et al.*, 2009; Minin & Suchard, 2008).

Rabies is a globally distributed zoonotic disease caused by a BEAST stranded negative sense RNA virus from the Lyssavirus genus. Though capable of infecting any mammal, given virus

variants are typically maintained in distinct species-specific cycles within the orders Carnivora and Chiroptera (Rupprecht *et al.*, 2002). The disease causes thousands of human deaths every year, predominantly in Asia and Africa where the virus circulates endemically in domestic dogs (*Canis lupus familiaris*) (Knobel *et al.*, 2005; Shwiff *et al.*, 2013). The vast majority of these deaths (~99%) are caused by bites from rabid dogs, instilling fear into the many communities that live under continuous threat from a disease that is almost invariably fatal but entirely preventable. Although the role of domestic dogs as key vectors of rabies is recognised, much less is known about the dog-associated RABV variant than wildlife variants such as raccoon or skunk RABVs circulating in North America (16). Moreover, while epidemic expansions of wildlife RABV have been well documented and studied (e.g. Biek *et al.*, 2007; Kuzmina *et al.*, 2013; Real *et al.*, 2005b) we know little about the persistence and spread of rabies in endemic landscapes.

Characterising the spatial scales of canine rabies dispersal is a critical step towards identifying the processes and factors driving its dynamics and the scale at which control strategies need to be implemented. On a global scale, canine RABV exhibits a strong phylogeographic structure with the distribution of seven distinct major clades reflecting the position of major barriers, such as oceans and mountain ranges, or historical mass human colonisation/migration events (Bourhy *et al.*, 2008; David *et al.*, 2007). However, it is unclear whether this genetic structure will persist in endemic scenarios or at smaller scales, and how much it is influenced by human-mediated dispersal. Indeed, on a regional scale this landscape structure becomes less distinct: some landscape features, for example, geopolitical boundaries can act as apparent barriers to movement, as seen in North Africa (Talbi *et al.*, 2010), whilst contradictory patterns of synchronous cycles of RABV across multiple countries (Hampson *et al.*, 2007) and repeated cross-border incursions (Hayman *et al.*, 2011) have also been observed elsewhere in the continent.

While there is a growing understanding of the epidemiology of canine rabies in Africa (Hampson *et al.*, 2007; Lembo *et al.*, 2008), effective rabies control is still hindered by limited knowledge of some of the key drivers of viral transmission and spread. Mass dog vaccination is the mainstay of successful rabies control but requires sustained coverage of at least 70% (Townsend *et al.*, 2013; WHO, 2005). In addition, spatial heterogeneity may affect how vaccine is most effectively distributed to interrupt key transmission corridors and target regions seeding RABV dispersal. An important aspect of this heterogeneity is the impact of human factors on RABV transmission, which has direct implications for control, including the design and scale of interventions necessary to interrupt transmission and maintain freedom from disease. For example, movement of people between urban and rural areas and the dog meat trade have been postulated as means of spreading RABV through human-mediated dog movements in rabies-endemic countries in Asia (Ahmed *et al.*, 2015; Denduangboripant *et al.*, 2005; Tao *et al.*, 2009). Uncovering the viral population structure and dynamics of RABV in Tanzania may identify similar patterns attributed to human-mediated movements that can

aid the identification of sources key to viral persistence.

There have been few in-depth spatial epidemiology studies of RABV in sub-Saharan Africa, probably owing to the lack of resources for effective surveillance including sample collection (Nel, 2013). However, recent studies provide intriguing insights into the dynamics of rabies in certain parts of Africa (see Mollentze *et al.* (2014b); Talbi *et al.* (2010)), indicating a degree of spatial structure with evidence of long distance movements facilitated by humans. RABV spatiotemporal dynamics in East Africa in particular are poorly resolved and very few sequences are publicly available. At present little is known about the genetic diversity or structure of RABV in Tanzania other than coarse phylogenetic analyses of partial or full nucleoprotein gene (N gene) sequences, limited to well studied regions (Kissi *et al.*, 1995; Lembo *et al.*, 2007). Many phylogenetic studies of rabies have focused on partial genome sequences, in particular the N & glycoprotein (G) genes as they have functions essential to viral propagation and pathogenesis (respectively). The N gene, although highly conserved, can provide enough genetic differentiation to characterise geographic lineages (e.g. Horton *et al.*, 2015; Kissi *et al.*, 1995; McElhinney *et al.*, 2011), whereas the the less conserved G gene has been used for more detailed epidemiological characterisation and host adaptation studies (Coetzee & Nel, 2007; Holmes *et al.*, 2002; Nadin-Davis *et al.*, 1999). Whole genome population studies of RABV have not yet been attempted despite their great potential to provide a better understanding of the processes determining rabies spread and persistence.

3.2.1 Phylodynamic inference

Phylodynamics has become an increasingly popular framework to enhance our understanding of infectious disease transmission dynamics and evolution. The field embraces the integration of additional data and models in phylogenetic reconstructions to extract information on the epidemiological and evolutionary processes underlying transmission (Grenfell *et al.*, 2004). Powerful Bayesian phylogeographic models have been used to shed light on spatial processes (mainly for viral pathogens), including reconstructing spatio-temporal patterns and determining the importance of specific factors influencing the spread of disease (e.g. Carvalho *et al.*, 2015; Lemey *et al.*, 2014; Lu *et al.*, 2014). Importantly, Bayesian methods provide a formal statistical procedure to model trait evolution, e.g. geographical locations, phenotype, using explicit spatial diffusion models while accounting for phylogenetic uncertainty (Lemey *et al.*, 2009, 2010; Pagel *et al.*, 2004). Box 3.1 provides an overview of phylodynamic concepts and definitions, including a graphical representation of spatial diffusion models.

Viral phylogeographic models treat spatial locations as an inherited viral trait, with the history of spatial processes captured in reconstructed phylogenetic trees (Faria *et al.*, 2011). Spatial locations can be classified as discrete (e.g. country of origin) or continuous distributions (geographical coordinates) with diffusion modelled using a *continuous time Markov chain (CTMC)*

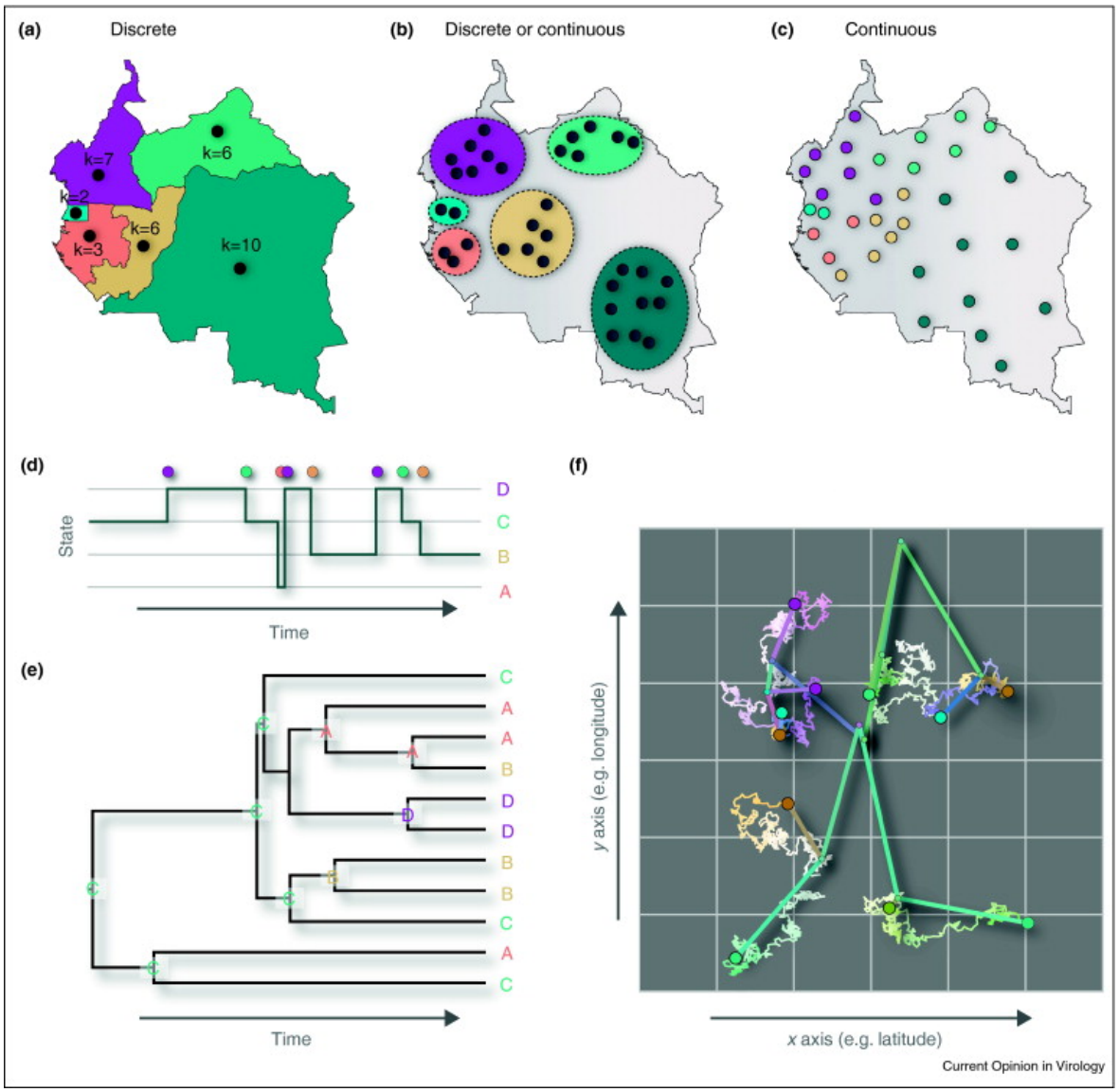


Figure 3.1: Hypothetical scenarios of sequenced samples' spatial distributions (top panels) and the modelling assumptions underlying discrete and continuous phylogeography approaches (bottom panels). Top: (a) recorded sample locations have a coarse geographic resolution requiring the distribution to be classified by discrete spatial units e.g. country of origin; (b) an intermediate scenario where more spatial detail is known but the distribution is still amenable to discretisation; (c) a continuous distribution of samples labelled with known geographical coordinates e.g. latitude,longitude. Bottom: (d) a graphical representation of a CTMC path for discrete phylogeography showing transitions between states through time for four discrete states A,B,C & D. Transitions from state i to state j are shown as jumps in the path and colour-labelled to indicate the end state j; (e) the CTMC process uses information provided by the observed trait data (at the tips of the tree) to model locations along each branch of the tree and infer the most probable ancestral states at internal nodes; (f) diffusion in continuous time and space is modelled using relaxed Brownian diffusion models to account for dispersal rate heterogeneity across the phylogeny. The panel shows an example of a Brownian diffusion process, where straight lines represent branches of a tree projected on a two-dimensional map and squiggly lines show the diffusion pathways from tips. Reprinted from Current Opinion in Virology (Faria *et al.*, 2011) with permission from Elsevier.

within a full probabilistic framework (Lemey *et al.*, 2009, 2010). The models use information provided by observed trait-associated sequence data, represented by the tips of a phylogenetic tree, to jointly estimate the phylogeny and infer the most probable spatial location of ancestral nodes (Faria *et al.*, 2011). Figure 3.1 provides a graphical representation of diffusion scenarios and model processes for discrete and continuous sampling distributions, explained in more detail below.

In discrete space, as explored in this chapter, the CTMC is characterised by a transition rate matrix that defines transitions between pairs of discrete locations (which can have asymmetric or symmetric rates) (Lemey *et al.*, 2009; Minin & Suchard, 2008). The history of transitions can be summarised by performing *Bayesian stochastic search variable selection (BSSVS)* allowing the application of *Bayes Factor (BF)* tests to identify important dispersal pathways and determine the most parsimonious explanation of the diffusion process (Lemey *et al.*, 2009). In addition, *robust counting* can be implemented to estimate the expected number of transitions between locations through time providing a quantitative measure of viral lineage migrations (Minin & Suchard, 2008). Such summary statistics provide information on viral origins, the importance of specific viral dispersal routes, source/sink dynamics and other spatial processes underlying transmission (Faria *et al.*, 2011).

In such cases where sampling schemes are not amenable to discretisation continuous diffusion can be modelled using a Bayesian implementation of multivariate *Brownian diffusion (BD)* models as an analogue to the transition model described above (Lemey *et al.*, 2010). Strict BD models assume homogeneous rates of diffusion through time across the phylogeny, an assumption that can be unrealistic and statistically inefficient in a viral diffusion scenario (Lemey *et al.*, 2010). However, *relaxed random walk (RRW)* models can be employed to increase model flexibility by accommodating variation in diffusion rates across branches in the phylogeny (Lemey *et al.*, 2010). Continuous diffusion models offer more realistic reconstruction over discrete phylogeographic models, fully exploring diffusion across 2-dimensional space rather than limiting to discrete locations (Faria *et al.*, 2011) and I explore these in 4.

In this chapter, finely resolved space-time-genetic data, including whole genome sequences, was used to determine the spatial and temporal dynamics of endemic RABV at both a regional and local ($<100\text{km}^2$) scale in Tanzania. Specifically, I aimed to 1) evaluate the utility of whole genome resolution to generate in-depth information from a small epidemiological window; and 2) characterise the dynamics of rabies virus in an endemic system including the role of human-mediated transport using a discrete phylogeography approach.

3.3 Materials and Methods

3.3.1 Samples

For this study, 59 new whole genome sequences were obtained from animal hosts (primarily domestic dogs) from 9 regions in Tanzania sampled between 2003-2012. A previously sequenced sample (RV2772; accession: KF155002 (Marston *et al.*, 2013)) from the same study area was included in the dataset and used as a reference sequence (see Table B.2). The main study area encompassed the Serengeti District in the northwest Tanzania where approximately half the samples (n=33) were obtained from active surveillance which enabled the collection of brain material from suspect rabid animals and GPS coordinates and dates recorded as described in Hampson *et al.* (2009). The remaining 27 samples were obtained opportunistically from other regions in Tanzania as part of active surveillance by the Tanzanian Ministry of Livestock and Fisheries Development and the Tanzanian Veterinary Laboratory Agency. All samples were sent to the Animal & Plant Health Agency (APHA) in Weybridge, UK, for processing.

In addition, 50 new partial N gene sequences (405bp) from Tanzania were obtained via reverse transcription-polymerase chain reaction (RT-PCR) and Sanger sequencing using samples archived at APHA (Table B.2).

3.3.2 RNA extraction and WGS

Total RNA was extracted at APHA directly from brain tissue using TRIzol, according to manufacturer's instructions (Invitrogen). Precipitated total RNA was re-suspended in molecular-grade water at a 1:10 dilution and quantified using a NanoDrops spectrophotometer (Thermo Scientific). Samples were sequenced on a range of next generation sequencing platforms during NGS protocol optimisation (see Appendix B text). The majority of samples (n=48) were sequenced by the following method: TRIzol-extracted viral RNA was depleted of host genomic DNA using the on-column DNase treatment in RNeasy plus mini kit (Qiagen) as per manufacturer's instructions with elution in 30µl molecular grade water. This was followed by host ribosomal RNA depletion using Terminator 5'-phosphate-dependent exonuclease (Epicentre Biotechnologies), as detailed in (Marston *et al.*, 2013). First and second strand cDNA was synthesised using a Roche cDNA synthesis system kit with random hexamers (Roche). Resultant cDNA was quantified using Picogreen dsDNA quantitation reagent (Invitrogen) and ~1ng of each sample used in a "tagmentation" reaction mix using a Nextera XT sample preparation kit (Illumina), according to the manufacturer's protocol minus the bead normalisation step. DNA libraries for each sample were quantified using a Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) or a Qubit assay kit (Life technologies), and average library size was measured with a high-sensitivity DNA Bioanalyzer chip on a model 2100 Bioanalyzer (Agilent).

Sample libraries were transported to the MRC Centre for Virus Research at the University of Glasgow, UK, for the final steps of library preparation and sequencing. Individual libraries were pooled and normalised to equimolar concentrations at a suitable plexity (x24 for MiSeq runs). Libraries were sequenced as 150bp paired-end reads on an Illumina MiSeq. Additional sequencing was conducted on a NextSeq 500 platform (Glasgow Polyomics at the University of Glasgow, Glasgow, UK) and reads merged with MiSeq reads to increase coverage for poorly sequenced samples (see Appendix B). Statistics for the overall NGS success rate of rabies virus samples processed in this thesis can be viewed in Appendix B, Table B.1.

3.3.3 Bioinformatics and sequence analysis

Raw reads were assessed in FastQC (Andrews, 2010) and Trimmomatic (Bolger *et al.*, 2014) was used to trim 3' ends, remove adapter contamination and to filter based on quality with default parameters. Filtered reads were mapped to the previously sequenced genome of Tanzanian RABV sample RV2772 with BWA mem version 0.7.10 (Li & Durbin, 2009) and converted to bam file format using SAMtools v. 0.1.18 (Li *et al.*, 2009).

A conservative SNP calling routine was implemented in GATK utilising the UnifiedGenotyper tool to identify high confidence SNPs, which were passed according to GATK filters on statistics for strand bias (FS>60, SOR>4), mapping quality (MQ<40, MQRankSum<(-)12.5), read position (ReadPosRankSum<(-)8.0) and depth of coverage (DP<5). Indels were filtered if FS>200 and ReadPosRankSum<(-)20.0, and further manually inspected for inclusion (e.g. dismissed if near a homopolymer run). Consensus sequences were built using a custom script in R where filtered SNPs were called with a 75% consensus rule (positions with <75% consensus were given a IUPAC code for the corresponding ambiguous base call) and genome positions with a depth of coverage less than one were labeled "N". Potential SNP calls that failed only the depth filter, that is, had a depth <5 but >1, were passed if the same polymorphism has been present as a high confidence SNP in at least two other samples. two other samples. Otherwise, the position was given an IUPAC code representing the population-level calls and the potential SNP. In addition, a set of consensus sequences using a more relaxed approach to SNP calling was produced which involved strict calls of all SNPs with depth>1 and gaps filled with the majority population consensus sequence. These relaxed consensus sequences were used to produce initial starting trees for BEAST analyses (see 3.3.6).

Sequencing resulted in 93-100% coverage of the genome, with >99% genome coverage achieved for 95%of samples and a median depth of coverage of 75 (range: 6-1871, see Table B.2). Nextera XT is a transposase-based method of library preparation and sequence reads typically miss the ends of the genome; however, as the ends of lyssaviruses are highly conserved (Kuzmin *et al.*, 2008; Marston *et al.*, 2007), it is unlikely that any informative variation was missed. I, therefore, consider my analyses to be based on genome-wide variation and henceforth refer

to my dataset as whole-genome sequences. Consensus sequences were aligned using MAFFT v7.149b (Katoh & Standley, 2013) and submitted to GenBank (accession numbers: KR906734-KR906792). 2.4

3.3.4 Phylogenetic reconstruction

Initial datasets of i) partial N gene 405bp (1317 sequences); and ii) full N gene 1350bp (674 sequences) sequences isolated in Africa were constructed using sequences retrieved from GenBank and including new Tanzanian isolates sequenced for this study (59 new WGS samples and 50 new partial genome sequences). Following maximum likelihood (ML) phylogenetic reconstruction with the initial sequence datasets, subset trees were extracted for samples in the Africa 1b clade (430 samples in the partial N dataset and 100 in the full N dataset).

Alignments for whole genome, full N and partial N gene were created in MAFFT (Katoh & Standley, 2013) and I estimated phylogenetic relationships using both ML and Bayesian methods. ML phylogenies were estimated in RAxML (Stamatakis *et al.*, 2012) with a general time reversible (GTR) nucleotide substitution model and a gamma distribution model of among-site rate variation. A Chinese dog RABV sequence from GenBank (Accession no: FJ712193) was used as an outgroup and node support was evaluated with 1000 bootstrap replicates. Bayesian phylogenetic reconstruction was conducted in BEAST v1.8.1 (Drummond *et al.*, 2012) using a posterior distribution of trees (without a molecular clock model). Phylogenies were visualised and annotated in R using the packages adegenet (Jombart, 2008) and APE (Paradis E., 2004) and maps were made in R with Maptools (Lewin-Koh *et al.*, 2012) and sp packages (Bivand *et al.*, 2008; Pebesma & Bivand, 2005). The degree of spatial admixture at large phylogeographic scales, i.e. sub-continental and country level, was quantified by an association index (AI) using BaTS software with beast phylogenies (Parker *et al.*, 2008).

3.3.5 Selecting an evolutionary model

An initial nucleotide substitution model was chosen based on the model selected by PartitionFinder (Lanfear *et al.*, 2012), an open-source program that selects the best-fit partitioning schemes and models of molecular evolution for nucleotide alignments. Whole genome alignments were partitioned into sets of nucleotides, one for each codon position (CP) in each gene (5 genes) and one for concatenated non-coding regions i.e. 16 sets in total were assessed in PartionFinder. Model scheme selection was based on the best AIC score from a greedy search of substitution models, which favoured a GTR model with partitioning into three CPs (CP123).

In addition, model tests by comparison of marginal likelihood estimates using path sampling

(PS) and stepping stone (SS) sampling implemented in BEAST (100 path steps and a chain length of 100,000 steps) were used to test varying levels of complexity in the substitution model. Non-coding sequence was concatenated and partitioned as a 'gene' with its own evolutionary model. Specifically I tested the HKY model with: 1) a gene-specific nucleotide model with gene-specific rate variation; 2) a gene-linked CP partitioned model with among codon position rate heterogeneity and homogeneous rates amongst genes; and 3) a gene-specific CP partitioned model with among codon position and among gene rate heterogeneity. Model types 2 & 3 were also tested with CP112 and CP123 partitioning schemes. Following the results for the best HKY model I did a final step comparing the most supported HKY model with the same structure but using a GTR model. Results (Table B.3) strongly favoured GTR and CP models (CP123 was best supported) but there was no support for partitions according to genes, which all had similar rates of substitution. This significantly reduced the complexity of the model and is an important finding in the context of analysing RABV whole genome sequence.

3.3.6 Bayesian evolutionary analyses

Bayesian Markov-Chain Monte Carlo (MCMC) analyses were performed using BEAST v1.8.1 (Drummond *et al.*, 2012) and the BEAGLE library (Ayres *et al.*, 2012). Based on model comparisons the most supported evolutionary model was a general time reversible model (GTR) with different substitution parameters for codon positions one, two and three (GTR123 + CP123 + Γ123) and homogeneous rates amongst genes, with a GTR + G substitution model for non-coding sequence. A relaxed molecular clock with a lognormal distribution was used to model rate variation among branches (molecular clock models were tested by PS and SS, Table B.4). A *Bayesian skyline* plot model with 10 groups was implemented as a flexible tree prior, which estimates the *effective population size* through time directly from the sampled nucleotide sequences while accounting for phylogenetic and coalescent uncertainty (Drummond *et al.*, 2005). To reconstruct the spatial dynamics of RABV spread in Tanzania I implemented a discretised diffusion process among 9 regional sampling locations as an asymmetric CTMC model (Lemey *et al.*, 2009). Three independent Markov-chain Monte Carlo chains with 50 million states and a sampling frequency of 50,000 were combined in LogCombiner after discarding at least 10 per cent burn. Posterior distributions were inspected in Tracer v1.6 (Rambaut & Drummond, 2014) to ensure adequate mixing and convergence. Initial analyses revealed issues with tree likelihood convergence. Therefore, a CTMC model using the relaxed version of consensus sequences, which contain fewer ambiguities, was implemented first and the maximum clade credibility (MCC) tree used as a starting tree for the final CTMC models with conservative consensus sequences.

To estimate the most significant pathways of viral dispersal between regions, a BSSVS procedure was implemented to identify the best supported diffusion rates through BF testing

(Bielejec *et al.*, 2011; Lemey *et al.*, 2009). For the per lineage rate of migration (Kühnert *et al.*, 2011), a conditional reference prior (Ferreira & Suchard, 2008) is commonly used but for this data resulted in convergence problems with some of the BSSVS parameters. Instead, we used an exponential prior with a mean of 0.01, which gave the most robust results out of a range of values tested (Table B.6).The degree of spatial admixture was scored using a modified Association Index (Lemey *et al.*, 2009; Wang *et al.*, 2001) and quantified using the inferred number of transitions between locations estimated by Markov jump counts (Minin & Suchard, 2008) along the branches of the posterior tree distribution (models without BSSVS used the conditional reference prior). A summarised history of Markov jump counts was used to identify movements between regions that occurred on very short branches and thus over unusually short time frames. Dogs rarely move 1km from their homestead (Hampson *et al.*, 2009; Woodroffe & Donnelly, 2011) and Hampson *et al.* (2009) found a maximum distance of 20km between linked RABV cases in the Serengeti District. Lineage migrations between regions, representing distances >100km, were therefore considered unlikely to be attributable to dog movement alone if they occurred over a period of 2 years or less and were instead interpreted as being the result of human-mediated movement. In addition, the same form of discretised diffusion model was used to assess diffusion at a broader scale, that is, between the North and South of Tanzania, with Pemba Island classed as a third discrete state. A BF test in the program SPREAD (Bielejec *et al.*, 2011) was used to identify well-supported migration pathways (BF>3). Sampled trees were summarised as an MCC tree with median node heights using TreeAnnotator v.1.8.1, and Figtree v1.4.2 was used to visualise trees and the inferred ancestral locations for internal branches.

3.4 Results

3.4.1 Geographic resolution: partial vs. full viral genomes

Consistent with previous large-scale phylogenetic studies, partial genome phylogenies indicated that RABV in sub-Saharan Africa falls into several regional groups with viruses from Eastern Africa generally being genetically distinct from those in western, central and southern parts of the continent (Figures B.1&B.2). Of the four major lineages of RABV in Africa (Bourhy *et al.*, 2008; David *et al.*, 2007; Talbi *et al.*, 2009), only the Cosmopolitan clade, and more specifically the Africa 1b lineage, was detected in Tanzania,

Table 3.1: Raw median genetic distance within each of the five main rabies virus lineages identified in Tanzania.

	Genetic distance	
RABV lineage	**No of substitutions per site**	**No of SNPs**
Tz1	9.47E-03	110
Tz2	2.27E-03	27
Tz3	8.83E-03	95
Tz4	4.17E-03	49.5
Tz5	0	0

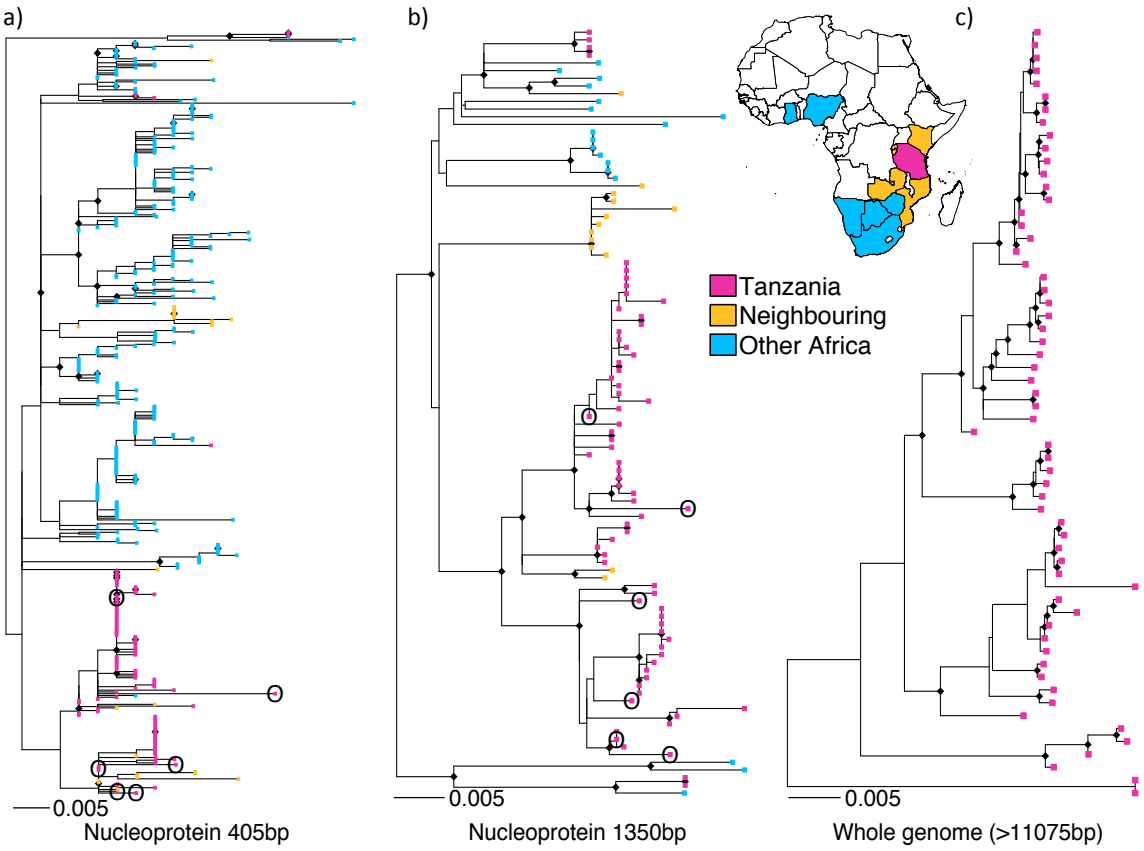


Figure 3.2: ML trees derived from datasets of rabies virus sequences from the Africa 1b clade for increasing levels of genome coverage: (a) 430 sequences from African countries highlighted on the map for a 405bp fragment of the nucleoprotein gene, (b) 100 sequences of full 1,350bp nucleoprotein gene from the same countries (except Botswana, Ghana, Kenya, and Zimbabwe); and (c) sixty full or near-full genome sequences (range: 11,076-11,923 bp) from Tanzania. Trees are scaled by number of substitutions per site and node symbols indicate nodes with bootstrap support ≥0.8. Historical samples from the Serengeti District (~20 years old) are circled in partial genome trees.

as found previously (Lembo *et al.*, 2007).

Within the Africa 1b lineage there was evidence of admixture between Tanzania and neighbouring countries and occasional long-range admixture at a continental scale (ML trees in Figure 3.3 & Bayesian maximum clade credibility trees in Figure B.3). Sequences clustering most closely with Tanzanian sequences came from Kenya, which shares a border to the north. While partial genome data was sufficient to identify such large-scale spatial patterns, these data did not provide adequate resolution to distinguish between samples at sub-national scales within Tanzania. The proportion of nodes with bootstrap support ≥80% was only 0.11 (Bayesian posterior probability ≥90%: 0.18) for partial N and 0.27 (Bayesian: 0.41) for full N gene phylogenies. Furthermore, out of the 60 Tanzanian WGS, 60% were identical at the partial N and 25% at the full N gene level. In contrast, maximum likelihood and Bayesian trees based on whole genome sequences were fully resolved and well supported (proportion of

nodes with ML bootstrap support ≥80%: 0.86; Bayesian posterior probability ≥90%: 0.89), even when samples had been taken in close spatial and temporal proximity. For example, RV2498 and RV2499 were sampled a day apart from the same village in Morogoro region: both samples only differed by a single SNP for the full N gene but were differentiated by 25 SNPs in the whole-genome alignment. This large divergence strongly suggests that the two samples are not from the same chain of transmission, which might have been the conclusion based on partial genome data. The median raw pairwise genetic distance between Tanzanian whole-genome sequences was 259 (range: 0-608) nucleotides and between the Serengeti District samples was 120 (range: 2-212) nucleotides, showing considerable diversity at even a small spatiotemporal scale. However, much of this high divergence is due to the presence of multiple lineages, some of which are evident even based on partial genome data. Using WGS, we identified five distinct lineages with high posterior probability support (annotated in Fig. 3.3)- median pairwise genetic distance within each lineage is listed in Table 3.1. Bayesian phylogenetic reconstruction of WGS yielded a mean evolutionary rate of 1.62×10^{-4} substitutions/site/year (95% highest posterior density [HPD] 5.51×10^{-5} to 2.61×10^{-4}), similar to previous estimates for N gene and G gene evolution (Ahmed *et al.*, 2015; David *et al.*, 2007; Talbi *et al.*, 2009).

Table 3.2: Degree of within-country spatial admixture in Tanzania measured using a modified Association Index (AI: 0 indicating complete population subdivision and 1 panmixis) for RABV sampled for this chapter and Algerian and Moroccan RABV sampled by Talbi *et al.* (2010). (BCI=Bayesian confidence interval)

Country	**AI 95% BCI**	***P*-value**	**No of sequences**	**No of locations**	**Median distance (km)**	**Min**	**Max**
Algeria	0.67 0.62-0.73]	<0.001	117	20	233.08	23.27	674.62
Morocco	0.55 [0.51-0.63]	<0.001	133	28	326.45	28.62	926.70
Tanzania	0.70 [0.60-0.79]	<0.001	60	9	439.76	39.05	1088.26

3.4.2 Phylogeography of RABV in Tanzania

Within Tanzania I found evidence of phylogeographic structure (AI=0.70,P<0.001), similar to other estimates of African intra-country AI values (see Table 3.2) (Talbi *et al.*, 2009). Compared with the strong spatial structure indicated between countries and larger spatial aggregations (Table 3.3) this indicates more fluid and dynamic dispersal patterns within Tanzania, as has been found in other African countries (Talbi *et al.*, 2010). Across the posterior distribution of trees there were 24 (95% HPD: 21-28) independent lineage movement events. Using a summarised history of Markov Jump counts across the phylogeny I found that 43% of these transitions occurred in the most recent ten years of the phylogeny (Fig. 3.3c). A

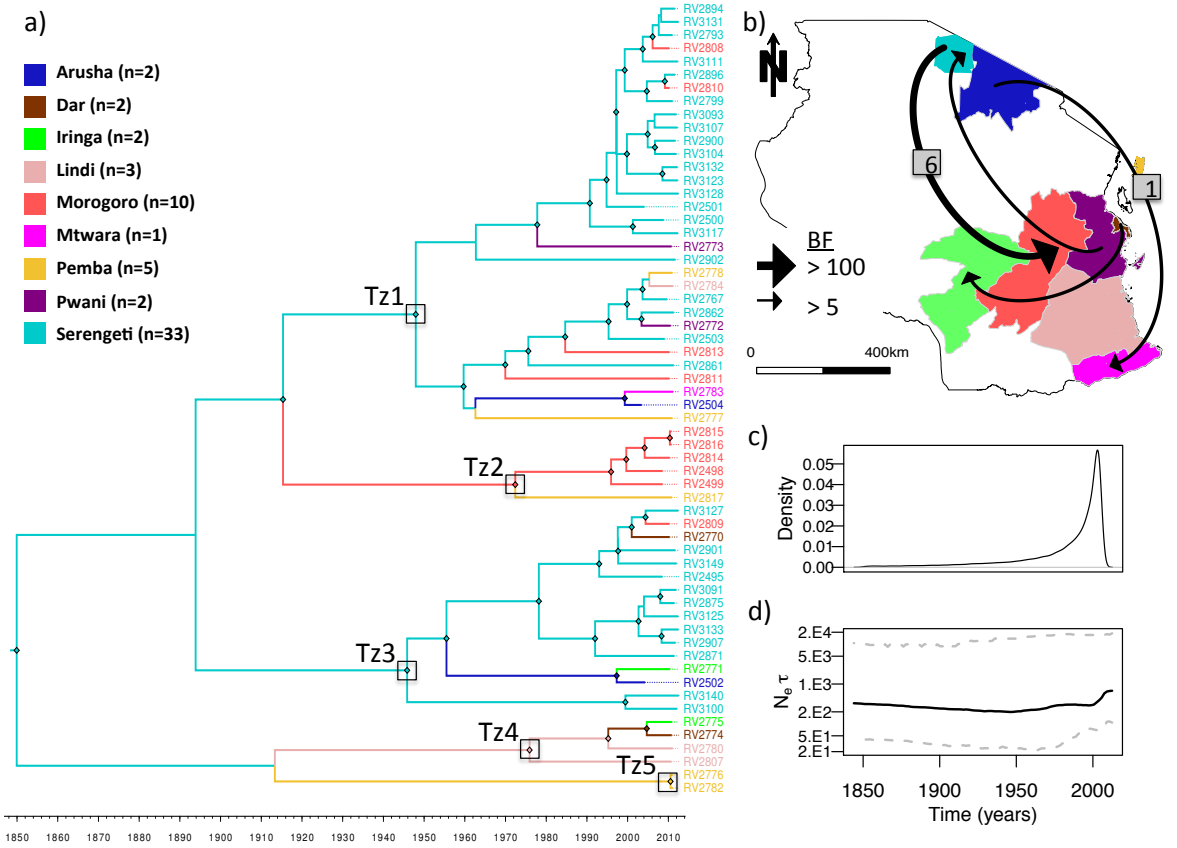


Figure 3.3: Regional phylogeography among sixty rabies virus whole-genome sequences sampled in Tanzania from 2003 to 2012: (a) an MCC tree with branches coloured according to the most probable posterior location of its descendent node inferred by discrete-state phylogeographic reconstruction in BEAST. Five major phylogenetic groups (Tz1-5) are annotated on the tree and node symbols indicate node posterior support $\geq$0.9. (b) The four most significant dispersal pathways indicated by BF results from a BSSVS procedure in BEAST with the median number of transitions estimated by Markov jump counts indicated in cases where posterior support for a transition was $>$0.7. (c) Markov jump densities for total number of transitions through time. (d) Bayesian Skyline plot showing N_{et}, the product of the effective population size (N_e), and the generation time (in years) through time.

BSSVS procedure in BEAST identified eighteen potential diffusion pathways to explain the observed phylogeographic patterns in the posterior distribution. However, only four received substantial support based on BF$>$5 (Fig. 3.3). Support was particularly strong for dispersal from the Serengeti District to Morogoro (BF=135.30), and Markov jump counts estimated a median of six (range: 3-10) migrations along this dispersal route, with at least one (range: 1-3) occurring on a branch representing a period of less than 2 years. Most lineages were sampled in more than one region in Tanzania, with some distributed across a larger geographic area than others (Fig. 3.4). The largest lineage, Tz1, contains not only a cluster of Serengeti samples but also encompassed samples from a larger geographic extent and was found in eight out of the nine districts sampled. The Bayesian skyline plot revealed that the effective

population size has remained fairly constant over the past 150 years (Fig. 3.3d). Because of the much higher availability of samples from the Serengeti District, I also conducted a coarser phylogeographic analysis grouping sequences into 'North', 'South', or 'Pemba Island'. This identified a predominance of north to south dispersal (estimated thirteen independent movements compared with one movement south to north) and evidence of dispersal back and forth between the North of mainland Tanzania and Pemba Island (Fig. B.4).

3.5 Discussion

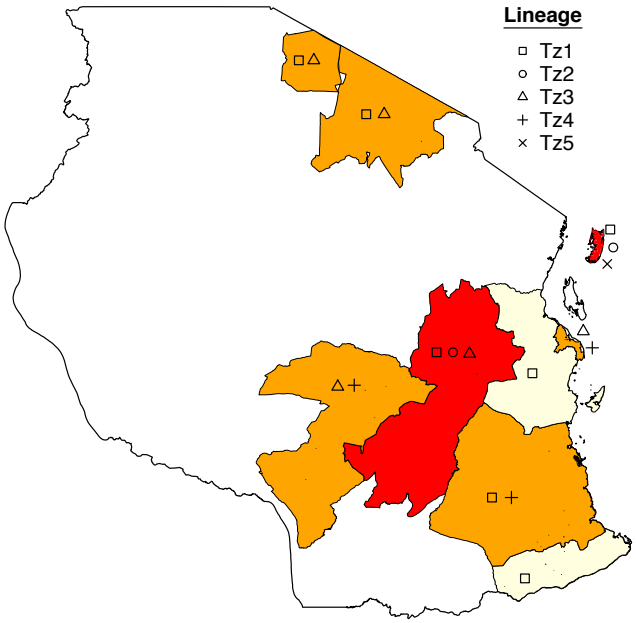


Figure 3.4: Spatial distribution of rabies virus lineages sampled from regions in Tanzania between 2003 and 2012 with a colour gradient (yellow to red) indicating the total number of lineages (low to high) sampled in each region.

Using viral genetic data from a hierarchy of spatial scales and varying levels of genome coverage I was able to demonstrate the advantages of whole genome resolution to describe the spatio-temporal dynamics of endemically circulating canine rabies viruses.

I found a clear phylogeographic structure between countries in Africa, which could be identified with partial genome data. This suggests that the majority of dispersal occurs at a within-country scale and control programs would be most appropriate at a national scale with strategies to deal with potential incursions from other countries. However, at regional and local scales within Tanzania the discriminatory power of partial genome data became too limited to reveal fine-scale population structure that could aid the effectiveness of control interventions. In contrast, WGS provided the resolution to genetically distinguish between all samples and produced a well-supported phylogeny. While sub-genomic information has utility at a coarse phylogeographic scale (e.g. Horton *et al.* (2015)) this finding supports the application of WGS for studies aiming to discern population structure at a scale most relevant to control.

Although large-scale population structure according to sub-continental areas (Figure B.1) or

country specific lineages (Figure B.2) was apparent from sub-genomic data I also found incidences of occasional large-scale admixture. I continued to find evidence at increasingly fine scales: Tanzania and other countries within the Africa 1b clade showed signs of admixture particularly when countries shared a border; and within country movements of RABV facilitated by humans were a feature of Tanzanian RABV. Even at a very local scale, within the Serengeti District, I found multiple co-circulating lineages. This recurrent theme may be a characteristic of endemic RABV in Africa reflecting decades of endemic circulation and human-mediated introductions. Similar patterns have been observed in canine rabies-endemic countries in South and Southeast Asia (Ahmed *et al.*, 2015; Lumlertdacha *et al.*, 2006; Matsumoto *et al.*, 2013), where the presence of co-circulating lineages was attributed to a combination of historical introductions from neighbouring countries and human-mediated movements.

Table 3.3: Degree of spatial admixture between rabies virus samples from Africa according to an Association Index (AI). Datasets of partial (N405) and full (N1350) nucleoprotein sequences were tested at two levels of spatial aggregation: 1) Sub-continent geographical partitions relative to Tanzania (3 states: Tanzania, neighbouring country, other African country); and 2) Country of origin. BCI, Bayesian confidence interval.

Data	**Spatial clustering level**	**Number of groups**	**AI [95%BCI]**	***P*-value**
N405	Sub-continent	3	0.04 [0.02-0.05]	<0.01
	Country	14	0.06 [0.05-0.08]	<0.01
N1350	Sub-continent	3	0.06 [0.04-0.08]	<0.01
	Country	10	0.11 [0.09-0.13]	<0.01

Much of the viral population structure I found within Tanzania is consistent with initial invasive waves that have persisted for long periods endemically, with structure eroding over time and aided by human-mediated movement. Historical accounts describe a rabies outbreak in southern regions of Tanzania in the mid 1950s, which spread throughout the country and was recorded in the Serengeti District in the late 1970s (Magembe, 1985; Siongok & Karama, 1985). This invasion perhaps shaped the initial RABV population structure in Tanzania. The timeline of regional scale migrations (Fig. 3.3c) indicates a strong rise in viral dispersal around this period, possibly reflecting increasing human connectivity. Samples in Tz1 (Figure 3.3) share partial N gene identity with samples from Kenya (Figure 3.2), indicating a shared evolutionary history that may relate to an initial outbreak across both countries. Lineages Tz1 and Tz3 appear to have a wide geographic distribution (Figure 3.4) and I found evidence of a general north to south dispersal pattern in Tanzania (Figure B.4). This highlights the potential for widespread dispersal should an incursion occur from Kenya and suggests a need for co-operative, cross-border rabies management once a national control program is established. Variation in the spread of different lineages across Tanzania may reflect the influence of heterogeneous landscape features or dog population structure in impeding or facilitating RABV dissemination. Identifying and

quantifying such features is the logical next step to provide additional information that can inform control programmes, such as identifying and strengthening pre-existing barriers. Recent extensions of phylogeographic techniques have highlighted how this might be achieved using an integrated approach combining evolutionary and ecological analyses to quantify drivers of viral transmission (Lemey *et al.*, 2014; Magee *et al.*, 2014).

The ancestry of Tz4 & 5 (denoting lineages found in the southern mainland and on the Island of Pemba, respectively) points toward an introduction from the north of Tanzania (posterior probability=0.62), but the uncertainty of this ancestral location and the ancestral node itself (posterior=0.45) suggest these may be distinct historical lineages with low-level persistence or undetected circulation elsewhere in Tanzania. Alternatively these clusters may represent instances where lineages from external sources have more recently invaded and resulted in sustained transmission in Tanzania- partial genome phylogenies indicate Tz4 is related to samples from countries south of Tanzania, South Africa or Mozambique. This again underlines the threat of re-invasion or introduction from external sources and the potential value of whole genome resolution to robustly and more accurately identify sources of new introductions that occur during control programmes. To date, the vast majority of RABV genetic data from Africa comes from partial genome analyses; however my data suggest that whole genome characterisation would be valuable and should be an aim for the future.

Islands represent isolated landscapes with natural barriers to dispersal, but incidents of RABV outbreaks instigated by human-mediated introductions (e.g. to islands in Indonesia (Susilawathi *et al.*, 2012; Windiyaningsih *et al.*, 2004)) have often been recorded. Sequences from the island of Pemba were suggestive of multiple introductions from various sources with evidence of dispersal to and from the mainland (Fig. B.4). Samples from Pemba were scattered throughout the phylogeny and the most divergent lineage Tz5 consisted of two samples from Pemba. The distribution of these lineages may reflect earlier invasions of RABV into Pemba from elsewhere in Tanzania (Tz1 and 2) and the African continent (Tz5), resulting from Pemba's location on an important trading route. However, since these lineages were not resampled and no cases have been detected from Pemba in over a year (Lushasi pers. comm) RABV may have been only transiently circulating. Nevertheless, lessons from Indonesia, where lack of a swift and coordinated response to a rabies incursion led to an epidemic (Townsend *et al.*, 2013), highlight the importance of active surveillance and rapid response to incursions.

Although I found evidence of RABV spatial structure within Tanzania, it was evident that dispersal was also frequent, with at least one long distance migration occurring within a small temporal window (<2 years). Lineage movements occurring on branches representing short evolutionary times, such as those identified between Serengeti and Morogoro (>750 km), indicate rates of dispersal much higher than those recorded for endemic wildlife rabies (Biek *et al.*, 2007) and imply human mediated movements as seen in parts of North Africa

(Talbi *et al.*, 2010). Movement of pastoralist and agro-pastoralist communities from the Lake zone in Northern Tanzania to southern regions, e.g. Morogoro, have been ongoing since the 1950s (Walsh, 2008), with a major influx reported from 2003-2006 (PINGO's Forum, 2013), attributed to climate change induced drought and forced evictions from newly protected areas (Kideghesho *et al.*, 2013). During these migrations, pastoralists are accompanied by their dogs, which may facilitate the long distance movement of animals incubating the virus prior to transmission. On further inspection, I found that Morogoro samples indicated as instances of long-distance translocation came from rural areas where pastoralists are likely to migrate to, while other Morogoro samples formed a distinct cluster (Figure 3.3). I found three RABV lineages in Morogoro from only ten samples (Figure 3.4). Quantifying networks of human-mediated movements (including livestock trade) would provide a valuable proxy for connectivity for many zoonotic diseases affecting domesticated animals.

Further to my findings of regional level admixture in Tanzania I also observed considerable diversity at a very local scale, that is, within the Serengeti District, with several lineages co-circulating. These lineages appear to have persisted for at least 20 years (older Serengeti samples obtained from GenBank also cluster within these lineages, indicated in Figure 3.2) despite dog vaccination campaigns having been conducted in the district since 1996. While vaccination coverage has varied substantially across years and between villages (Viana *et al.*, 2015) rabies incidence has at times significantly declined e.g. falling by 97% in the late 1990s (Cleaveland *et al.*, 2003). Yet the genetic data shows that, despite substantial declines in incidence, these lineages must have persisted at very low levels within the district or subsequently reinvaded from neighbouring districts. Without sampling the surrounding districts it is not possible to distinguish between these possibilities. The skyline plot indicated a stable effective RABV population size (N_e) through time (Figure 3.3). N_e represents the effective genetic diversity of the virus at the host population level and can be used as evidence of temporal trends such as viral population expansion or a response to selective pressures or control efforts (Hall *et al.*, 2016). While a stable N_e could be expected for an endemic pathogen, it is worth noting that I found no evidence of viral population size reductions in response to vaccination campaigns. Rabies control efforts across much of Tanzania have been very limited until recently and high turnover in the dog population (Hampson *et al.*, 2009) likely contributes to the stable persistence of rabies in Tanzania. It has also been noted that geographic structure can obscure local fluctuations in subpopulations while maintaining the appearance of a constant N_e in skyline plots (Carrington *et al.*, 2005) and sampling schemes can be problematic in skyline estimates (Hall *et al.*, 2016).

This chapter represents a snapshot of RABV dynamics in Tanzania, indicating that human movements have disseminated RABV out of locally endemic areas at scales relevant to control, that is, administrative units such as regions or districts. These frequent translocations have probably lead to the existence of multiple co-circulating lineages (Fig. 3.4), but relatively few introductions lead to sustained chains of transmission that are detectable as lineage movement

events. However, in disease systems closely associated with human activities, the probability of successful translocation and establishment is likely to be much greater. This suggests that without some level of regional co-operation, Tanzania will be unable to eliminate rabies and maintain freedom from disease. Human movements are often in response to social drivers, which could be used as signals for increased vigilance and surveillance in at risk areas.

My findings highlight the use of WGS to uncover fine scale transmission patterns that can directly inform control efforts. However, sub-genomic approaches can still have utility at a coarser scale and are more easily obtained, particularly when sample quality is an issue. In particular, they can be used to initially identify admixture between countries, which may indicate the necessity of coordinated regional control programs and surveillance. Co-circulation of multiple lineages and introductions facilitated by humans appear to be a feature of endemic rabies virus and complicate the design of a sustainable control strategy. However, using whole-genome data, we were able to identify sources of dispersal within Tanzania that could direct efforts toward surveillance and control. The finding that humans play an important role in the dynamics of RABV in Tanzania suggests that increasing awareness and dog vaccination in "high-risk" communities such as pastoralists could help to reduce long-range dispersal. Moreover, the design of enhanced surveillance and containment strategies to mitigate human-mediated incursions and maintain disease freedom should be a priority once control programs are established and elimination is being targeted.

CHAPTER 4

Quantifying the effects of landscape heterogeneity on the local-scale phylodynamics of an endemic zoonotic virus.

4.1 Abstract

Heterogeneities in the landscape shape the distribution, abundance and movements of host and pathogen populations, influencing host-pathogen interactions and ultimately disease transmission. Quantifying the extent to which the distribution of viral diversity is influenced by landscape processes is crucial to successful intervention. Landscape features including natural physical barriers and anthropogenic features have previously been found to influence the spread of rabies and increasingly the contribution of human mobility to disease spread proves to be an important determinant of viral dispersion in general. In my study system, where the disease is endemic and the main host is inherently tied to human behaviour and ecology, naturally restrictive landscape features may well be circumvented by human mediated dispersal/connectivity. Various landscape features including topographic variables, population measures and vaccination data were represented as resistance surfaces and integrated into phylogeographic models of spatial diffusion. Their efficacy as predictors of diffusion was determined by means of comparison to a null isolation by distance model using various summary statistics of the diffusion process, including the number of inferred migrations between landscape-informed spatial clusters and generalised linear model results. Isolation by distance patterns accounted for much of the variation in phylodynamic diffusion estimates but landscape predictors had discernible effects. Importantly, there was quantitative evidence to support vaccination as a control measure for canine rabies. Support for other landscape features including rivers as barriers to rabies dispersal and road networks as a predictor of connectivity across the landscape was also found. This heterogeneity could be exploited to create more effective control measures such as the stratification of vaccination effort.

4.2 Introduction

Infectious diseases are of significant concern to animal and human health across the globe, with the highest burden often placed on low-income countries. While many control measures have successfully interrupted transmission and drastically reduced incidence in some cases e.g. polio, measles, guinea worm, achieving elimination is a long and difficult road (Klepac *et al.*, 2013). Landscape heterogeneity and connectivity play a key role in transmission and pathogen persistence at multiple spatio-temporal scales and continue to be important factors in the latter stages of elimination (Klepac *et al.*, 2013). Identifying and quantifying key landscape features and processes is therefore central to successful intervention.

Landscape heterogeneity inherently shapes the distribution, abundance and movements of host and pathogen populations, influencing host-pathogen interactions and ultimately disease transmission. The "landscape" occupied by any pathogen is a spatially complex environment defined by the contribution of many ecological, physical, and socio-cultural features, which

makes the identification of key predictors particularly challenging. The field of landscape epidemiology aims to gain a comprehensive understanding of the features and processes that impact the incidence or risk of disease by taking an integrated approach, utilising different types of data to illuminate patterns of circulation (Brunker *et al.*, 2012; Ostfeld *et al.*, 2005). With the increasing accessibility of so-called "big data" (Kao *et al.*, 2014; Pfeiffer & Stevens, 2015), including whole-genome characterisation of pathogen populations, integration is increasingly desired to advance spatial and temporal analyses but remains challenging.

A promising direction for integration lies within a phylogeography framework, which traditionally uses genetic information to establish relationships between historical processes and contemporary geographic distributions but has recently been extended to allow the inclusion of a generalised linear model (GLM) parameterisation to simultaneously determine the impact of potential predictors on diffusion (Lemey *et al.*, 2014, 2009). Recent examples of such an approach include the identification of global live swine trade as a major driver of swine influenza A virus diffusion (Nelson *et al.*, 2015), and the discovery that the major factors driving dispersal differ according to serotype in foot and mouth disease virus in South America (Carvalho *et al.*, 2015). These approaches have typically been applied to large scale scenarios of epidemic spread but here I assess their utility in an endemic context at a local scale (henceforth referring to an administrative district scale with a spatial area $<100\text{km}^2$). In practice this would help identify dispersal pathways for targeted interventions, whilst pre-existing structure may be used to inform the roll out of control measures, taking advantage of barriers for the placement of vaccination campaigns.

In this chapter I compare a combination of recently developed and novel phylogeographic approaches to determine the contribution of different landscape predictors on the spatial spread of canine rabies virus (RABV) in a local endemic system in Tanzania. As introduced in Chapter 3 phylogeographic models can consider diffusion under a discrete or continuous scenario- here I exploit both methods to incorporate landscape heterogeneity. Discrete phylogeography has become a well established framework to model diffusion between defined areas given the strength of summary statistics enabled by robust counting, Bayesian stochastic search variable selection (BSSVS) (both utilised in Chapter 3) and the GLM approach mentioned above. As part of this Chapter it also provides the opportunity to determine at what spatial resolution the impact of landscape heterogeneity can be detected.

Continuous phylogeography offers a more realistic model of diffusion for many study systems but it has rarely been used to look at local scale dynamics or endemic situations (Raghwani *et al.*, 2011; Trewby *et al.*, 2016). The method produces a full spatial dispersal history based on sampled sequences which can be used to characterise the rate, direction and variation of spatial spread through time (Lemey *et al.*, 2010). Landscape heterogeneity is a key source of spatial variation in viral dispersal but there is no formal framework to incorporate and quantify landscape variables in phylogeographic models. Dellicour *et al.* (2016) recently de-

veloped a statistical approach based on assigning a landscape variable "weight" to lineage dispersal paths extracted from continuous phylogeographic reconstructions, which measures the association between landscape variables and lineage movement. As part of this Chapter I directly incorporate landscape heterogeneity into reconstructions by rescaling the observed spatial location data according to landscape-informed resistance distances and explore the effect on dispersal estimates.

RABV is particularly susceptible to landscape influences as it requires direct contact between infectious and susceptible hosts to transmit but we have yet to gain a better quantitative understanding of the landscape characteristics that influence transmission at local scales and in endemic scenarios. As with many zoonoses RABV transmission is affected by anthropogenic landscapes, particularly due to the inherent connection between dog and human populations. Naturally restrictive landscape features such as large rivers may be overcome in time through human mediated dispersal, resulting in the erosion of initial invasive structure. Contemporary patterns of RABV distribution in North Africa were found to reflect human-mediated dispersal, with road distances uncovered as the best observed predictor of diffusion (Talbi *et al.*, 2010). Major physical barriers such as rivers and mountain ranges have been indicated as barriers to RABV diffusion in wildlife systems (Rees *et al.*, 2008; Smith *et al.*, 2002) and global canine RABV dissemination (Bourhy *et al.*, 2008) but it has yet to be established if this type of feature acts as a barrier on a smaller scale. Importantly, deliberate modification of the landscape i.e. control interventions can also be considered in a landscape approach, potentially providing a measure of the success of interventions such as vaccination or highlighting particular features that can be exploited in a targeted campaign e.g. pre-existing barriers can be reinforced with a cordon sanitaire.

The study area focused on here covers the extent of the Serengeti District in Tanzania, where RABV has been circulating endemically since the 1970s and mass dog vaccination campaigns have been undertaken for the last decade. The area has been well studied since vaccination campaigns began and a unique dataset of genetic, epidemiological and landscape data is available, including per village vaccination coverage and dog density. This provides an exciting opportunity to characterise the landscape processes influencing the spread and persistence of RABV in a local endemic scenario. The results could have direct implications for control efforts in the area and should provide a generalisable framework to quantify landscape attributes in other host-pathogen systems.

4.3 Materials and Methods

4.3.1 Sequence data

Brain samples were obtained from rabid animals in the Serengeti District of northwest Tanzania between 2004-2013, along with a record of the precise GPS location and date symptoms started for each case. Samples were processed at the Animal & Plant Health Agency in Weybridge (APHA) as described in Chapter3. The same conservative SNP calling protocol was also implemented to produce consensus sequences. In total, I compiled a dataset of 152 whole genome sequences, consisting of 119 new sequences and 33 sequences previously utilised in Chapter 3. Sample details, including epidemiological data and sequence statistics can be seen in Appendix C.

4.3.2 Landscape and predictors

The study landscape was defined as a spatial grid encompassing the Serengeti District (spatial extent: xmin=637638.2, ymin=9757825.5, xmax=705238.2, ymax=9835425.5) with a resolution of 100 x 100m cells. Potential landscape features were characterised as surface models by assigning cell values to represent the assumed facilitating or impeding impact of a predictor on RABV diffusion. Each landscape feature was considered individually and therefore no standardisation of cost values was required. For example, rivers have previously been identified as barriers to RABV dispersal and cells containing a river were therefore assigned a high resistance value. Landscape features assumed to facilitate diffusion were given resistance values according to the reciprocal of their assumed conductance value e.g. roads were assigned an arbitrary conductance factor of 1000 giving a resistance value of 0.001. Cells with no landscape heterogeneity were given a resistance value of one to represent uniform landscape and a null model of isolation by distance (IBD) was created, where all cell values were set to one. None of these cost values represent absolute barriers to diffusion. Pearson correlations between cost surfaces were calculated and can be seen in Table C.2.

Circuitscape (Shah & McRae, 2008) was used to generate a matrix of pairwise resistance distances between all RABV sample locations for each landscape resistance surface. The program uses a combination of circuit and graph theory to model connectivity according to the effective resistance between pairs of points (see McRae *et al.* (2008) for a detailed review). Landscape rasters are converted to graphs with each cell represented by a node and connections by undirected weighted edges. Resistance (i.e. edge weights) between two nodes was calculated as the average per-cell resistance value. An advantage to circuit theory methodology is that multiple connections between nodes can be considered (in this analysis 8 neighbours were considered for each node) accounting for the effect of multiple pathways

connecting points (McRae & Beier, 2007). The effective resistance distance is then computed as the value of a single resistor required to produce the equivalent resistance observed along all paths between pairs of points (McRae & Beier, 2007).

Details of the different landscape predictors tested and their assigned resistance values are indicated in Table 4.1 and final resistance landscapes are shown in Fig.4.1. Landscape data was sourced from the National Bureau of statistics in Tanzania and census data collected as part of an ongoing Wellcome Trust funded project (095787/Z/11/Z). Resistance surfaces for each predictor were formatted as follows:

4.3.2.1 Dog density

Dog density estimates were taken from a household census conducted in the Serengeti District in 2014-2015, which provided the GPS location and dog count for each household. The density point pattern was smoothed across the default raster grid using an isotropic Gaussian smoothing kernel with sigma=500 using the R package spatstat (Baddeley & Turner, 2005) with cell values expressing the estimated intensity values. As dog density is assumed to facilitate RABV diffusion, the reciprocal of these values was used as a resistance value in each cell. Serengeti National Park (SNP) areas were assigned a low resistance value, equivalent to 1 dog per km^2, to reflect the low dog density in this area and cells outside the district (~12%) were assigned random values from the Serengeti density data.

4.3.2.2 Elevation and slope

A digital elevation model covering the landscape was converted to raster format and elevation values used directly as resistance values. Slope values were calculated from the digital elevation model using the SDMTools package in R (VanDerWal *et al.*, 2014). All slope values were increased by a value of 1 to ensure comparison to a null IBD surface was possible. The resolution of the digital elevation model was not consistent across the extent of the landscape as some areas have been mapped in more detail, which resulted in some finer grained areas in raster grids (see B Fig.4.1).

4.3.2.3 Human to dog ratio (HDR)

Household census data was used to estimate the human to dog ratio per village, which was assigned directly as a resistance value in each cell.

Table 4.1: Details of landscape predictors and their assumed influence on rabies virus diffusion.

Predictor	**Hypothesised effect on dispersal**	**Cost value**	**Rationale**
Uniform (IBD) landscape	Na	1	Null model for isolation by distance testing spatial spread in ideal dispersal habitat
Dog density	Conductor	Inverse of actual values	Domestic dogs are the main maintenance host
Average vaccination coverage	Resistor	Inverse of actual values with no coverage assigned 1	Mass dog vaccination is the mainstay of effective rabies control and has been shown to significantly reduce incidence in the Serengeti in the past (Cleaveland *et al.*, 2003).
Elevation	Resistor	Actual values	Mountain ranges have been implicated as barriers to wildlife and canine RABV dispersal at large spatial scales (Bourhy *et al.*, 2008; Wheeler & Waller, 2008). There are no major mountain ranges in the Serengeti landscape (elevational range: 1196-1549m) but smaller changes in steepness may still influence diffusion.
Human to dog ratio	Resistor	Actual values	In areas with a higher human to dog ratio dogs with rabies are more likely to be caught and killed before transmitting.
Rivers	Resistor	1000	Evidence that rivers act as barriers to wildlife RABV dispersal in northeastern American and Europe and canine RABV at a global scale (Russell *et al.*, 2004; Smith *et al.*, 2002)
Major roads	Conductor	0.001	Human transportation networks have been implicated in the distribution of canine RABV (Denduangboripant *et al.*, 2005; Talbi *et al.*, 2010; Tenzin *et al.*, 2010)
Slope	Resistor	Actual values (+1)	The steepness of the landscape may provide a better characterization (than elevation) of the effects of hills and mountains on dispersal.
Susceptibles	Conductor	Inverse of actual values	A resistance surface to incorporate the effect of vaccination on the susceptible dog population.

4.3.2.4 Major roads and rivers

Shapefiles of major roads and rivers in the study area were converted to raster grids (one for roads, one for rivers) with the defined spatial extent and resolution. Cells containing a road feature were assigned a low resistance value of 0.001 to represent increased diffusion along roads relative to uniform landscape, all other cells were assigned a value of 1. River cells were assigned high resistance values of 1000 to reflect their influence as a barrier to diffusion.

4.3.2.5 Average vaccination coverage

Mass dog vaccination campaigns have been undertaken in the Serengeti District since 2002, with varying annual coverage across villages. I used an average annual % vaccination coverage per village across the 11-year period from 2002 to 2013 and assigned values to grid cells as resistance values [range: 6.43-100]. Rabies appears to have been locally eliminated in the SNP and there is a requirement for all dogs to be vaccinated within the park boundaries, therefore SNP cells were assigned the highest resistance cost of 100. Cells out-with the district and SNP were assigned the minimum observed average vaccination coverage of 6.43% as there is no formal vaccination initiative in these areas and therefore coverage is assumed to be low.

4.3.2.6 Susceptibles

The dog density estimates were depleted according to vaccination coverage to produce a cost surface representative of the susceptible host population.

4.3.3 Empirical tree distribution

To overcome the computationally intensive task of exploring phylogenetic tree space repeatedly in each set of analyses I first estimated a posterior distribution of trees inferred solely from sequence data. Sequence evolution was modelled as described in Chapter 3 with sequences partitioned according to coding and non-coding concatenated sequence and coding sequence further partitioned into codon positions 1+2+3. However, a simpler HKY substitution model was used as there were issues achieving convergence and a reliable tree set under a GTR model. This was implemented with a relaxed molecular clock and a Bayesian skyline model used as a flexible tree prior (Drummond *et al.*, 2005). MCMC analyses were performed using BEAST v1.8.1 (Drummond *et al.*, 2012) and the BEAGLE library (Ayres *et al.*, 2012). Five independent MCMC chain were run for 50 million steps, sampled every 50,000$^{\text{th}}$ and combined with LogCombiner v1.8.1. The combined posterior tree distribution was subsampled to

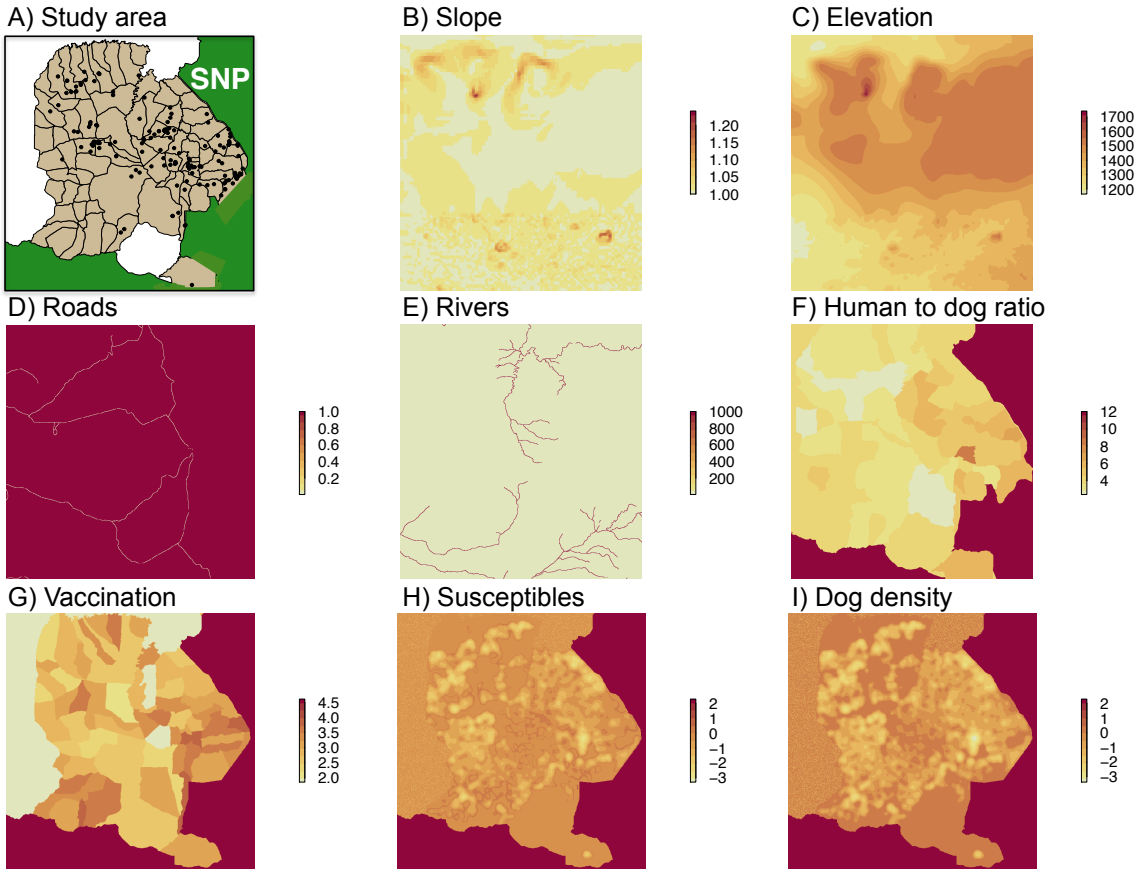


Figure 4.1: The study area used for analysis (A) showing the distribution of RABV cases sampled from villages in the Serengeti District (black circles) adjacent to the Serengeti National Park (SNP). Landscape resistance surfaces (B-I) are shown for individual landscape features with colours displaying increasing cost values (yellow to red). [Note cost values for G-I are log transformed for better visualisation.]

a set of 1000 trees to provide an adequate sample of phylogenetic uncertainty. Chains were inspected for stationarity and adequate mixing in Tracer v1.6 (Rambaut & Drummond, 2014) and a 10% burnin discarded from each. The resulting empirical tree set was used in all subsequent diffusion analyses to approximate phylogenetic uncertainty with the implementation of a transition kernel to randomly sample from the tree distribution (Pagel *et al.*, 2004).

4.3.4 Measuring diffusion in predictor-modified landscapes

4.3.4.1 Finding clusters for discrete diffusion models

Multidimensional scaling (MDS) in R was used to project cases in 2-dimensional space representing each landscape predictor in Table 4.1. MDS is a means of representing objects (in this case, rabid animal cases) as points in an N-dimensional space given a similarity or dis-

similarity matrix (in this case, a resistance distance). Here the aim is to produce a spatial configuration of RABV cases in 2 dimensions to represent the perceived proximity between cases according to landscape resistance. Rescaled coordinates from the MDS landscape projection can be used in phylodynamic reconstructions to explore the features influencing viral diffusion. Figure 4.2 shows an example of rescaled coordinates using resistance distances according to average vaccination coverage (see Fig.4.1G cost surface). In this example k-means clustering was used to assign five clusters according to the rescaled coordinates and these clusters were used as traits in a discrete phylogeographic reconstruction.

Each landscape feature was described by a matrix of resistance distances among pairs of RABV cases calculated by Circuitscape and MDS reproduced a rescaled configuration from the resistance matrix.Taking rivers as an example, which I consider as a barrier to RABV dispersal, cases separated by a river would have a larger resistance distance and therefore would be projected further apart in MDS space. MDS projected data were then partitioned according to varying levels of spatial aggregation using a k-means algorithm. The following methods were used to assess an optimum range of k-values to consider for discretisation:
1) Elbow method: the point of maximum curvature in a plot of number of clusters versus within-group sum of squares;
2) Partitioning around medoids: using the optimum average silhouette to estimate the number of clusters, R package: fpc (Hennig, 2014));
3) Model-based clustering: chooses the optimal model and number of clusters according to Bayesian Information Criterion for expectation-maximisation, R package:mclust(Fraley & Raftery, 2002));
4) Affinity propagation: a clustering algorithm that takes a pairwise similarity matrix and simultaneously considers all data points as potential cluster centres. The algorithm finds an optimum set of clusters that maximises the total similarity between data points and their cluster centres by an iterative process (Frey & Dueck, 2007), R package: apcluster (Bodenhofer *et al.*, 2011);
5) Gap statistic: a statistical procedure to formalise the "elbow" method by comparing the change in within-cluster dispersion to a reference null distribution (Tibshirani *et al.*, 2001), R package: cluster (Maechler *et al.*, 2015);
6) R package NbClust: provides 30 indices to determine the number of clusters (Charrad *et al.*, 2014).
Resulting spatial clusters (for each k in the optimum range) were used to specify the location state for each viral sequence in a discrete phylogeographic analysis (Lemey *et al.*, 2009). Diffusion between discrete locations was modelled using a non-reversible continuous-time Markov chain (CTMC) process, which uses a K x K infinitesimal rate matrix Λ of location change among K-discrete locations. The expected number of location state transitions in the ancestral history given the observed tip data was estimated using Markov jump (MJ) counts (Minin & Suchard, 2008) and a modified Association Index scored the degree of spatial admixture (Lemey *et al.*, 2009; Wang *et al.*, 2001). MCMC chains with a pre-defined tree space (the em-

pirical tree set) were run for 5 million steps and sampled every 500 to summarise parameters related to the diffusion process. Two measures were used to assess diffusion amongst clusters in comparison to a null model (i.e. diffusion in a uniform landscape):
(1) Migrations between clusters: a reduction in MJ counts (while keeping the number of clusters constant) across the phylogeny indicates that less dispersal is required to reconstruct the observed spatial pattern
(2) Degree of phylogenetic clustering: measure using a modified association index (AI) (Lemey *et al.*, 2009; Wang *et al.*, 2001), which reports the posterior distribution of association values relative to those obtained by randomising the tip locations. Low AI values represent strong phylogeny-trait association.
In this analysis fewer MJ counts and stronger phylogenetic clustering than expected under a null model is indicative of an informative predictor.

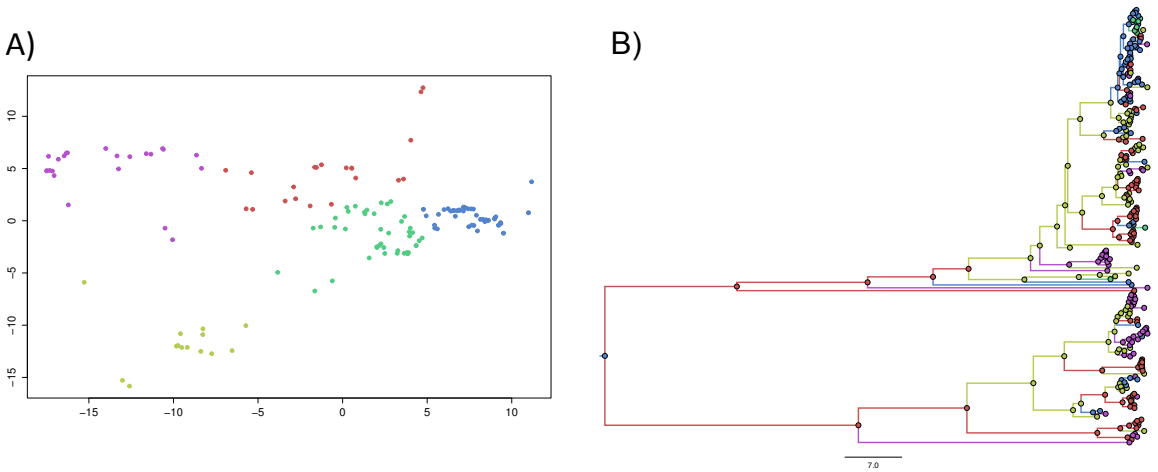


Figure 4.2: A) Multidimensional scaling in 2-dimensions to rescale the actual geographic locations of RABV cases in the Serengeti District according to average vaccination coverage resistance distances and consequent k-means clustering with k=5 (clusters coloured); B) Discrete phylogeographic reconstruction using k-clusters as traits.

4.3.4.2 Continuous diffusion

Spatial diffusion across contiguous landscape was modelled under the continuous phylogeography framework described by (Lemey *et al.*, 2010). Relaxed random walk models were implemented to allow dispersal rates to vary along branches according to Gamma or Log-normal prior distributions. To test the effect of different predictors in a continuous diffusion process Bayesian multidimensional scaling (BMDS) (Oh & Raftery, 2001) was used within the BEAST framework to project cases in 2-dimensional "landscape space" using resistance distances. A measure of the variation in spatial spread (coefficient of variation) across the phylogeny was used to compare the dispersal process in predictor-modified landscapes. A reduction the coefficient of variation suggests a dampened diffusion process in a given landscape

space. MCMC chains were run for 50 million steps with sampling every 50,000.

4.3.5 Testing the effects of landscape heterogeneity on diffusion

GLM diffusion parameterisation (Lemey *et al.*, 2014) of the discrete diffusion model was applied to estimate the influence of potential predictors on diffusion between discrete locations. Cases were partitioned into k-discrete locations by MDS as explained above using a euclidean distance matrix. This is slightly different to the above approach as the landscape is not manipulated before partitioning. Landscape predictors for the GLM model were constructed using the pairwise resistance distances between the centroids of each cluster and were log-transformed and standardised before their incorporation in the GLM. Pearson correlations between predictors were calculated (see Table C.3). In cases where the correlation was greater than or equal to 0.9 a GLM was also tested with one of the correlated predictors removed to ensure it had no effect on the results obtained.
In the GLM approach the migration rate matrix used to model diffusion is parameterised by a log linear function to incorporate a set of predictors on a log such that:

$$\log\Lambda_{ij} = \beta_1\delta_1\log(p_{1ij}) + \beta_2\delta_2\log(p_{2ij}) + ... + \beta_n\delta_n\log(p_{nij})$$

The relative contribution of each predictor p to the GLM can be measured via a coefficient β and a binary indicator δ determines the inclusion or exclusion of an individual predictor in the model. The indicator variables are estimated using BSSVS, which estimates the posterior probability of all possible models including or excluding each predictor and so results in an estimate of the posterior inclusion probability for each predictor. A Bernoulli prior probability distribution was used for δ to place an equal probability of inclusion or exclusion on each predictor (Lemey *et al.*, 2014). Bayes factors (BF) were calculated using δ estimates (Lemey *et al.*, 2014) to assess the level of evidence against the null hypothesis i.e. the observed predictor inclusion (pp_p) versus the prior opinion for predictor inclusion (qp_p):

$$\mathrm{BF}_p = \frac{\mathrm{pp}_p}{1-\mathrm{pp}_p} / \frac{\mathrm{qp}_p}{1-\mathrm{qp}_p}$$

A BF ≥ 3.0 was considered as the threshold for sufficient support against the null hypothesis, which corresponds to pp_p being 3-times more likely that pq_p (when a predictor is included 50% of the time). MCMC chains were run for 5 million steps and sampled every 500.

4.3.6 Overall evidence

Each landscape variable was ranked according to the strength of evidence to support their influence as a predictor of diffusion. Each set of results from the three phylodynamic methods was summarised as follows:

i) Results for predictor structured space were condensed to the larger spatial scales tested, k3-k6, as this appeared to be the most relevant spatial scale to test landscape effects. Each predictor was ranked based on the sum of the mean number of migrations and the mean AI value across these scales in relation to IBD effects.
ii) Continuous phylogeography results were ranked by maximum coefficient of variation values in descending order.
iii) GLM results were ranked according to Bayes Factor results; Only predictors with positive evidence in relation to IBD were considered and an overall ranking was calculated via a sum of the individual rankings with the lowest sum given the highest overall ranking. A penalty of 9 was awarded for each predictor that had no significant BF results.
Overall rankings were determined by the ascending order of the sum of individual rankings (and penalties), i.e. the smallest sum was ranked first.

4.4 Results

4.4.1 Diffusion in predictor-structured space

Based on tests to assess the optimum spatial discretisation for each predictor I found a range of values was necessary to account for uncertainty. Therefore discrete diffusion was tested between K-clusters ranging from 3 to 14 for each landscape predictor. In all case clusters structured according to IBD or landscape predictors had less dispersal and more phylogenetic structure than randomised data (results not shown). Heatmaps displaying measures of diffusion and phylogenetic structure for predictor-clusters relative to IBD-clusters are displayed in Fig.4.3. Colour ramps reflect estimates after IBD results have been subtracted to indicate support for a predictor relative to IBD alone. Actual values for the mean number of lineage migration events with 95% highest posterior density (HPD) intervals are also shown in each cell. Overall, there is a high number of migrations events, which may reflect the relatively fluid dispersal dynamic in this area. In each heatmap green cells indicate scenarios when there is support for a predictor relative to IBD. In Fig.4.3A green cells represent instances when less overall diffusion was observed i.e. structuring according to a predictor has reduced the number of migration between clusters required to explain the observed spatial pattern. Results varied according to cluster size but most predictors were consistently better than IBD at larger spatial scales (k3-k6). However, HPD intervals for each predictor overlap with those observed for randomised clusters (not shown) and IBD-clusters, therefore the null hypothesis that landscape resistance has no effect on diffusion cannot be dismissed.

As an additional measure I calculated an association index to assess the degree of phylogenetic structure according to each spatial discretisation. Low AIs imply a strong association between spatial and phylogenetic relationships.Results are summarised in Fig.4.3B, with green cells

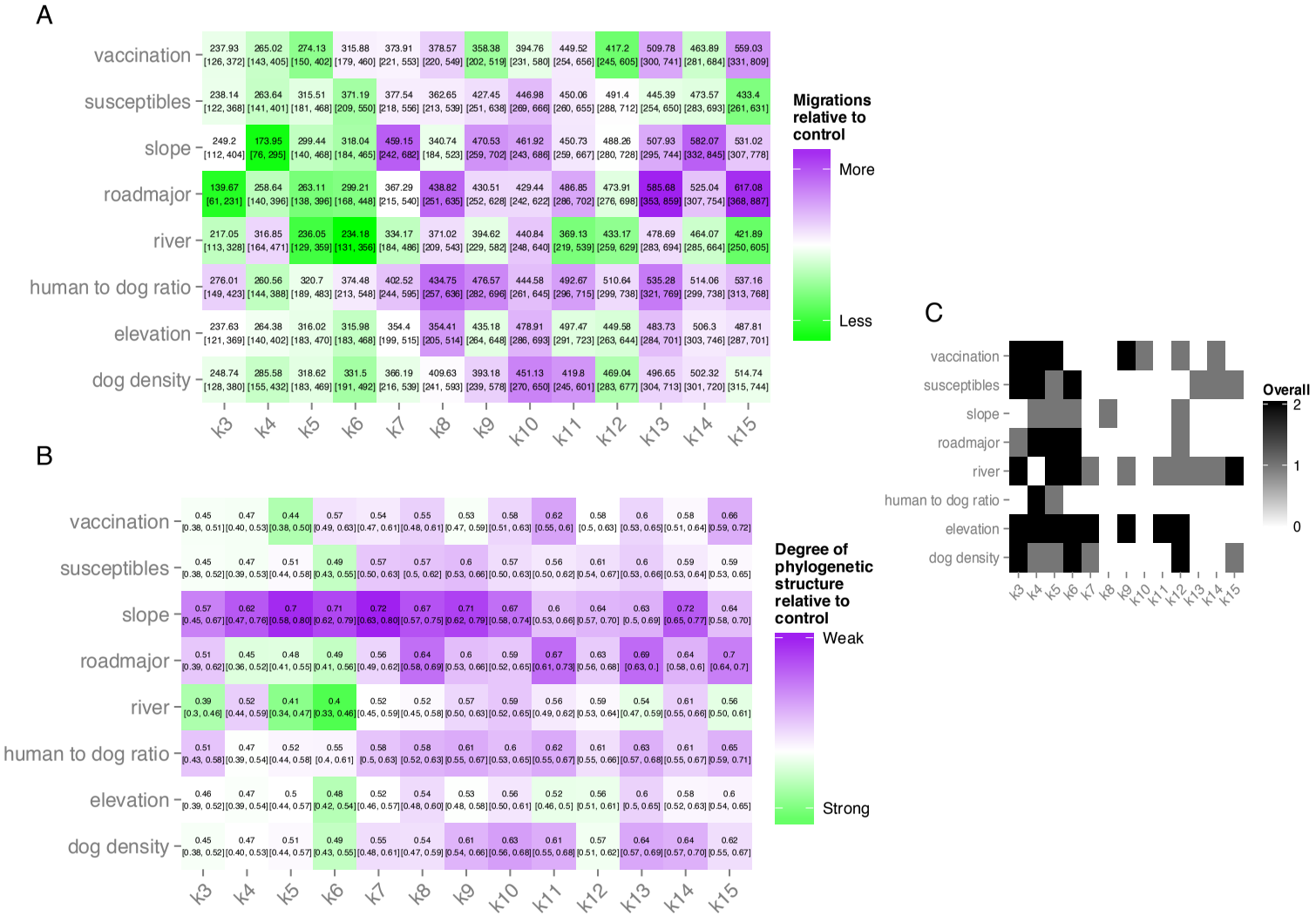


Figure 4.3: Heatmaps showing measures from the phylogeographic reconstruction of RABV spatial spread in landscapes modified according to different landscape features or processes in the Serengeti District. A) shows counts of the numbers of viral lineage migrations between k-discretised locations in each predictor-modified landscape and B) shows a measure of phylogenetic structure according to the same spatial discretisation via an association index value ranging from 0 (indicating complete population subdivision) to 1 (complete panmixis). Colour ramps in each heatmap represent values relative to a null isolation by distance landscape model, with green cells representing instances when results favoured the predictor over the null model, i.e. fewer migrations and lower ai values. C) a summary heatmap representing the overall support for each predictor

reflecting a higher degree of spatial structure than IBD and actual AI values with 95% HPD intervals are shown. All green cells represent instances when some support was found for a predictor relative to IBD, with bright green indicating strong combined support from both measures. Purple cells show instances when predictors had less support than IBD as explanatory variables. Phylogenetic structure tended to be stronger when there was a large reduction in the number of lineage migrations but results for the two measures were not always consistent, see Fig.4.3C. As suggested by individual heatmaps, diffusion was best explained at larger spatial scales, with strong support for rivers and roads in particular.

4.4.2 Landscape effects on continuous diffusion

Continuous diffusion in predictor-modified landscapes was estimated under lognormal and gamma RRW models. Modification of landscape space inherently modifies the spatial scale of diffusion by structuring data according to dissimilarity distances. The resistance distances used to modify landscape space for each predictor are not relative to one another and therefore comparison of diffusion rates could not be used to assess predictive power.

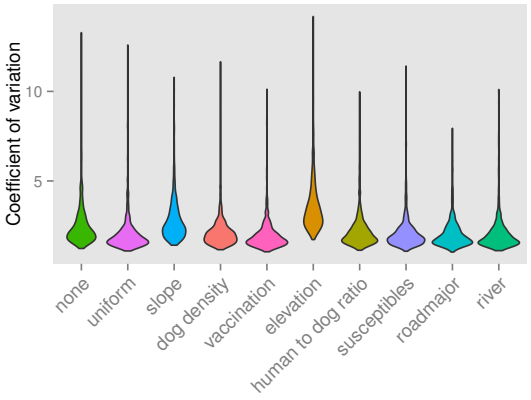


Figure 4.4: Variation in the diffusion of endemic RABV diffusion in landscapes modified according to spatial heterogeneity in different landscape features. Violin plots show the full posterior distribution of the diffusion coefficient of variation among lineages with width corresponding to the probability density of the data at each coefficient value.

A lognormal-RRW provided a better fit than a gamma-RRW mode, which accommodates a higher level of variation in branch-specific diffusion rates demonstrating the highly variable nature of RABV's spatial spread. The mean rate of spread in unmodified landscape was 5.75 km/year (95%HPD: 3.87-7.78), similar to estimates for enzootic wildlife RABV spread (Biek *et al.*, 2007; Lemey *et al.*, 2010) but around five times lower than estimates for dog RABV spread in North Africa (Talbi *et al.*, 2010). I assessed predictors in comparison to diffusion in uniform-landscape, which accounts for isolation by distance (IBD) in ideal dispersal habitat. The coefficient of variation (CV) in the diffusion rate coefficient among branches was used to quantify variability in the diffusion process, with a reduction relative to IBD indicative of a dampened diffusion process. This measure is robust to interpretation between different predictors as it measures dispersal relative to the mean and therefore is not influenced by the modified spatial scale of diffusion. Variation among branches remained relatively consistent under the influence of different predictors, but no-

tably the extent of extreme variation was reduced under nearly all models with landscape heterogeneity (except elevation). This effect was strongest for roads which had a maximum lognormal CV value of 7.96 compared to 12.62 for IBD. This suggests that landscape structure has accounted for some of the more extreme diffusion events observed when only IBD is considered.

4.4.3 Relative influence of predictors on diffusion

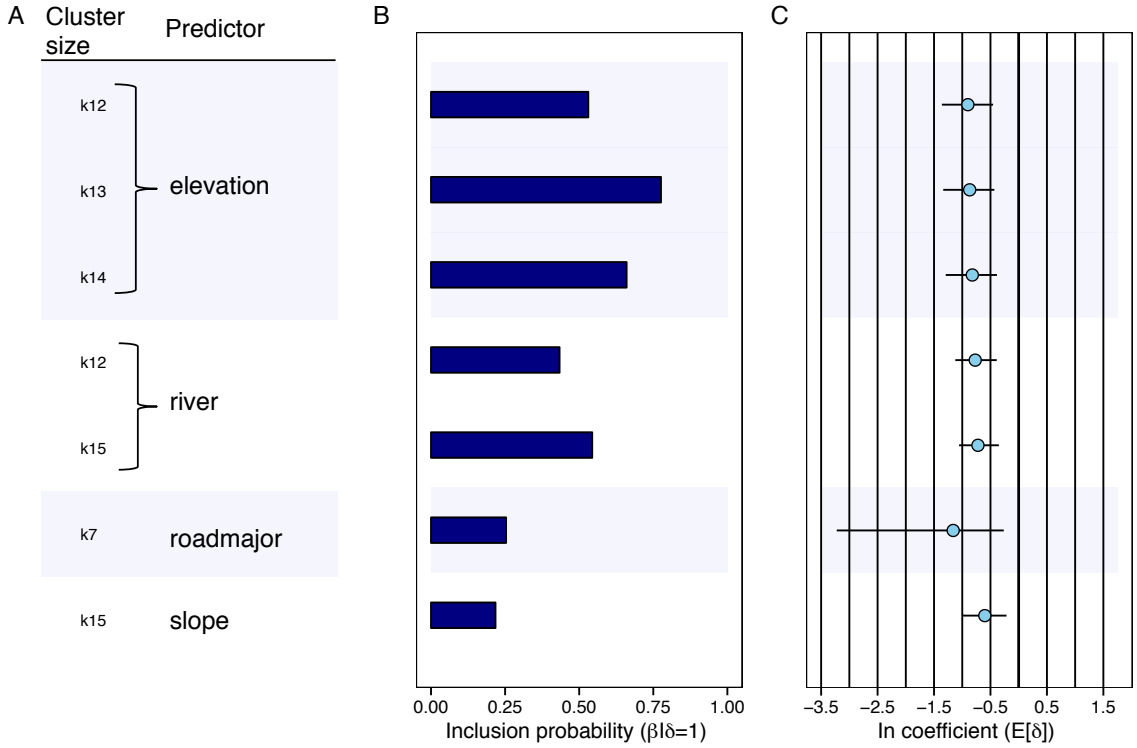


Figure 4.5: The support and contribution for predictors of RABV diffusion with Bayes Factor support >3 among k-discretised clusters in the Serengeti District: A) Predictors with the k-discretisation level at which they had significant effects on dispersal (e.g.. k7 corresponds 7 spatial clusters); B) support for each predictor represented by an inclusion probability (E[δ] and C) the relative contribution of each predictor indicated for log scale GLM coefficients (β) conditional on the predictor being included in the model.

I used a GLM approach within a Bayesian framework to identify landscape predictors driving the spatial spread of RABV in the Serengeti District. I also tested this approach under a range of spatial discretisations to ensure the effect of scale was accounted for. The majority of predictors did not yield significant results at any spatial scale, specifically dog density, susceptibles, vaccination and hdr had no discernible support. Results for predictors that reached a BF threshold of 3 are presented in Fig.4.5, which shows posterior inclusions probabilities and conditional effect sizes for each predictor. In instances where predictors were highly correlated (Table C.3, a simplified GLM model with the removal of one of the predictors was

performed and had no effect on BF significance levels.

In general, significant effects were found at smaller spatial scales i.e. with a higher level of discretisation. All predictors with significant support had a negative effect size, consistent with lower rates of diffusion as the effective resistance of the predictor increased. Note that predictors were proposed to be conductors or resistors of diffusion *a priori*, which were used to inform effective resistance calculations and hence effect sizes must be interpreted with this in mind. This means for an *a priori* conductor, e.g. roads which had a significant GLM result, negative effect sizes reflect lower rates of diffusion when road resistances were high i.e. when fewer roads were present in the landscape. Elevation was well supported (BF>10) at three spatial discretisation (k=12,13 &14) with an estimated negative effect size between -0.82 to -0.9 (on a log scale), indicating that viral lineage movement rates were lower at high elevations. Rivers also had reasonable support at two spatial scales (k=12 &15), again with a negative effect size indicating lower rates at higher levels of river resistance. Roads and slope each had marginally significant results (BF~3) at one spatial scale.

4.4.4 Overall support for landscape predictors

The overall support for individual landscape predictors is shown in Table 4.2.

Table 4.2: Overall support for individual landscape features as predictors of rabies virus diffusion in the Serengeti District. Predictors are ranked in terms of the strength of evidence relative to an isolation by distance landscape for each measure of diffusion applied in three different phylodynamic models. Discrete and continuous diffusion models tested the effect of modifying the landscape according to each predictor and the GLM approach tested the relative contribution of each predictor to the diffusion process in an unmodified, discretized landscape.

		Discrete		**Continuous**	**GLM**
Predictor	**Overall ranking**	Markov jumps	Association index	Coefficient of Variation	Bayes Factor
Rivers	1	1	1	3	2
Roads	2	2	2	1	3
Vaccination	3	4	4	2	9
Slope	4	3	8	5	4
Elevation	5	7	5	8	1
Susceptibles	6	5	3	6	9
Dog density	7	6	6	7	9
Human to dog ratio	8	8	7	4	9

4.5 Discussion

In this chapter I aimed to quantify the contribution of different aspects of landscape heterogeneity on the spatial spread of RABV at a local scale. I assessed three different phylogeographic approaches to identify and quantify the effect of different predictors on RABV diffusion, whilst also considering the inherent impact of scale on my results. Overall, there was some supporting evidence for all predictors in at least one of the phylodynamic approaches used but the strongest, most consistent support was found for rivers and roads. The effect of rivers and roads was best seen at larger scales in accordance with the natural delineation of the landscape created by this type of predictor. Roads as facilitators of RABV spread implies a strong human influence on diffusion, as also found for canine RABV in North Africa (Talbi *et al.*, 2010). In the continuous phylodynamic model roads had the strongest effect on the diffusion process, reducing variation in diffusion and potentially explaining more of the extreme diffusion events that might be expected via human-mediated movement. However, roads may simply best reflect the true accessibility of the landscape as they circumvent physical barriers and uninhabited areas. Either way they represent a possible route of RABV dissemination into and within the Serengeti District. Settlements on particularly busy routes could be subject to increased surveillance and public awareness and targeted with increased vaccine effort to interrupt spread. Alternatively, the effect of roads may actually be driven by surveillance bias if rabid dogs are more likely to be detected and sampled near roads. If this were the case some evidence to support human to dog ratios as a predictor might be expected but was not observed, since human population centres tend to exist in proximity to major roads.

Rivers have previously been found to reduce the dispersal of wildlife rabies in northeast America (Rees *et al.*, 2009; Wheeler & Waller, 2008) and Europe (Bourhy *et al.*, 1999) but my results suggest that they also have an influence on a much smaller magnitude and in an endemic system with more human interference. Further analysis considering how these two features interact on the landscape, e.g. bridges crossing rivers increases barrier permeability, would provide a more realistic measure of the status of rivers as barriers. As potential barriers to diffusion they could be exploited to delineate the landscape for optimal deployment of vaccine effort. For example, mathematical models determined that the best placement of vaccine corridors for RABV control in an outbreak scenario is behind a barrier to limit spread (Russell *et al.*, 2006). In an endemic situation where RABV is already present on both sides of the river (but for which I still found evidence of a barrier effect) the barrier could be best utilised through effective timing of vaccination campaigns on either side of the river or directing limited resources to less protected areas.

Importantly, I found quantitative evidence for a reduction in RABV dispersal due to vaccination coverage. This provides direct empirical evidence demonstrating the success of local interventions, namely the effectiveness of mass dog vaccination campaigns in the Serengeti

District. That I was able to find support is encouraging considering the low vaccination coverage estimates (range 6-48%) recovered from an 11-year period. While WHO recommends vaccination coverage should be ≥70% (WHO, 2013), this suggests that lower levels of coverage can also be successful in controlling canine RABV. This may be due to particular characteristics of the Serengeti landscape and therefore it is important that other influences on dispersal are identified. In addition, analyses incorporating temporal fluctuations in coverage (discussed below) and a resolution similar to dog density estimates obtained from census data (which may be possible in future) may provide more insight.

There was little evidence to support dog density as a predictor, which substantiates the paradigm that RABV transmission is not density dependent and population reduction as a control method will be ineffective (Hampson *et al.*, 2009; Morters *et al.*, 2013; Townsend *et al.*, 2013). Although highly correlated, using susceptible population density, which accounted for the depletion of susceptible hosts by vaccination, was slightly better as a predictor. Although supporting evidence is limited this also corroborates that vaccination has a discernible impact on RABV spread and should be continued as a control measure in the Serengeti District.

Spatial scale played an important role in the interpretation of results, even within my local landscape. This effect was tested by discretising cases according to resistance distances for each predictor under a range of spatial aggregations. Since the effect of predictors may vary according to the spatial scale under consideration I attempted to find an optimum discretisation level for each landscape predictor. However, there was large uncertainty surrounding estimations of the optimum number of clusters for each predictor, which led to a range of cluster numbers being defined as the test range for all predictors. The high number of migrations inferred between clusters in general (see numbers in Fig. 4.3) suggests that discretisation is not entirely appropriate to describe this local scale dynamic, indicating a more fluid diffusion process. Difficulties associated with geographic partitioning in phylodynamic models have previously been noted (Lemey *et al.*, 2014) but the consistency of my results across a number of similar spatial aggregations implies that the observed effects on diffusion are robust to specific partitioning effects such as cluster size. I found this to be true for a number of predictors, which showed consistently strong results at larger spatial aggregations but gradually diminished effects at higher resolutions. At higher resolutions the spatial scale will begin to approach the scale of local transmission among hosts in their natural home range and the landscape becomes increasingly more homogeneous, reducing the opportunity for spatial structure to arise. According to my results landscape structure in the Serengeti becomes less discernible at a cluster size ~k6 or k7 (see Fig. 4.3C). I also note that the Serengeti District is not a particularly heterogeneous environment and many landscape variables, particularly linear features such as rivers or roads, can only impact diffusion at relatively large scales. For example, rivers (see landscape structure in Fig.4.1) can only partition cases to a certain extent before further segregation becomes meaningless. Therefore choosing an appropriate partitioning scheme is an important consideration when implementing a discrete

analysis. While I chose to use an objective k-means algorithm to spatially discretise data, it may be more appropriate to structure data subjectively when there is a clear delineation of the landscape.

Conversely to the landscape-structured discretisation approach, I only found significant results at higher spatial resolutions under the GLM-diffusion approach which used resistance distances between geographic-cluster centroids as predictors in a GLM model. By using centroid positions to represent clusters we inherently lose a lot of information regarding the spatial structure of cases that may be informative in the overall diffusion process. At larger scales this is particularly problematic as the true spatial configuration of cases is oversimplified to the point that it no longer accurately reflects the original structure. At higher resolutions i.e. using a large number of centroids, the simplified spatial structure more accurately reflects the observed configuration and patterns are more likely to emerge. I found significant support for elevation (BF=12 to 38) and rivers (BF= 8 to 13) as predictors of diffusion at the higher end of the spatial resolution tested (k$\geq$12). Increasing the number of centroids might help to uncover more significant effects but becomes more computationally intensive as it increases the dimensionality of the CTMC rate matrix size and the number of parameterisations.

In contrast to the other approaches the GLM model assesses the relative contribution of predictors to the diffusion dynamic. Highly correlated resistance distances such as those for dog density and susceptible density (see Table C.3 present a problem in this analysis as they may explain the same variation. Simplified GLMs were performed to verify the results obtained with the full inclusion of all predictors. However, as shown by Talbi *et al.* (2010), resistances may be correlated but one or the other can still offer a marginally better fit as an explanatory variable. For example, here the distinction that susceptibles provides better explanatory power than density fits with expectations regarding the effect of vaccination. As many of the predictors tested have some correlation due to the underlying isolation by distance structure such subtleties may be important to best characterise the landscape dynamic under the simplest scenario. Results from alternative analyses such as the discrete phylodynamic model used in this Chapter could be used to evaluate correlated predictors prior to inclusion in GLM parameterisations.

Ideally, diffusion at the scale presented in this Chapter would be best explained under a continuous dynamic but at present there is no "tried and tested" method to incorporate landscape heterogeneity into continuous phylogeographic analyses. I integrated landscape heterogeneity using BMDS, morphing the spatial landscape over which a relaxed Brownian process was applied to model diffusion within a phylogeographic framework. The rational behind this approach was that diffusion in landscapes modified to incorporate the underlying heterogeneity would become less variable in modified landscapes with strong explanatory power. While this application shows promise there is no intuitive means of interpreting results and quantifying the influence of different predictors on diffusion parameters. In particular,

the application of BMDS to modify the spatial structure also modifies the possible spatial extent of diffusion and renders the direct comparison of diffusion rates impossible. If an appropriate normalisation procedure could be developed it may be possible to use diffusion rates as another measure of changes in the diffusion dynamic. The coefficient of variation of the diffusion rate was used to interpret the regularity of the diffusion process, with lower values indicating a more homogenised diffusion dynamic but there appears to be little power in the comparison of this statistic amongst landscape predictors. I found no clear distinction between the 95% HPD intervals of predictors and null models but noted a reduction in the most extreme variation for most predictors compared to an IBD model. The uncertainty surrounding the explanatory power of the landscape predictors (in all analyses) may stem from the consequence that most of the diffusion dynamic can be explained by IBD, leaving a very small proportion of "leftover" spatial variation to determine. There is a clear need to evaluate these problems using simulation models.

An important consideration in the overall approach used here is the use of resistance surfaces to represent landscape features or processes. Determining appropriate resistance values to represent different types of landscape feature is a common challenge in landscape ecology and there is currently no consensus on the best method to optimise resistance surfaces (Beier *et al.*, 2008, 2011; Spear *et al.*, 2010; Zeller *et al.*, 2012). Ideally, resistances should be parametrised according to relevant empirical data but often relies on expert opinion when such data are unavailable (Beier *et al.*, 2008). I parameterised resistance surfaces subjectively using arbitrary values to reflect the assumed resistance of linear features such as roads and rivers and assumed linear relationships between continuous variables and resistance. In the case of continuous variables such as elevation or vaccination, it may be more appropriate to consider non-linear relationships e.g. examining critical thresholds and classifying resistance values accordingly (Spear *et al.*, 2010). In addition, landscape predictors were tested with an *a priori* assumption on their effect as either a conductor or resistor but it is conceivable that some predictors may act conversely to their presumed effect. For example, roads, as mentioned above, could also be considered as barriers of diffusion if surveillance bias near roads facilitates earlier detection and removal of rabid animals.

Although my parameterisation of resistance values may not be ideal, they still scale with biologically meaningful quantities and reflect the relative effects of features on diffusion compared to uniform landscape, which is more important than the choice of absolute resistance values (McRae, 2006). However, some landscape processes may not be well represented by resistance surfaces, particularly when a temporal aspect to the heterogeneity is involved. Vaccination coverage in my analysis was summarised over an eleven-year window discarding potentially important temporal fluctuations, therefore the resultant resistance surface may not be informative enough to test its effect on diffusion processes. A recently phylodynamic application to relax the time-homogeneity assumption in phylogeographic reconstructions suggests a method could be developed to incorporate some of this temporal heterogeneity (Bielejec *et al.*, 2014).

The approach involves defining time-discretised "epochs" to which different infinitesimal rate matrices are applied and was applied in a phylogeographic context to test the effect of seasonality on the spatial dispersal dynamics of influenza H2N2, yielding a better model fit than a time-homogenous model (Bielejec *et al.*, 2014).

The analysis presented here mainly involved tested predictors independently (except the GLM approach) but a more biologically realistic approach would be to produce a multivariate surface to represent the different processes underlying diffusion. This creates a new set of considerations, including identifying collinearity between features and how to attribute relative resistance values to each predictor.

In summary, I assessed a number of methods to incorporate landscape heterogeneity into phylogeographic reconstructions. I found evidence supporting the impact of a number of predictors on RABV diffusion and the spatial scale at which these affects apply. There are a number of important considerations including the construction of resistance surfaces and ways to robustly interpret and quantify results that need to be addressed before these analyses can be applied directly in a meaningful way, for example to inform control efforts. However, I have demonstrated the potential for a general application of such techniques to identify important landscape features and processes driving the dispersal of disease and set a basis for further investigation.

CHAPTER 5

Inferring the dynamics of endemic canine rabies virus using high resolution space-time-genetic data

5.1 Abstract

Despite their enormous burden on global animal and human health, we still understand little about the dynamics of endemic pathogens. Rabies virus is a widespread zoonotic pathogen spread predominantly to humans by domestic dogs. Although entirely preventable through mass dog vaccination campaigns it still inflicts devastation on communities in Asia and Africa where it is endemic in dogs. Surprisingly little is known about endemic transmission dynamics and how rabies manages to persist despite a low basic reproductive number (R_0). Crucial information on transmission biology is hard to obtain as transmission itself is rarely observed. Here I use a subset of highly resolved dataset for animal rabies cases in the Serengeti District of Tanzania to exploit cutting-edge statistical techniques to incorporate all available data in an integrated Bayesian inference approach. This included integrating whole genome sequence data with contact tracing information and detailed spatio-temporal incidence data. Building on an existing Bayesian framework transmission trees were reconstructed from space-time-genetic data to show the local scale dynamic of endemic rabies virus. Specifically, the model dealt with how to incorporate genetic and contact data from only a proportion of cases and allow multiple origins of infection (exogenous to observed data or from an observed source). Tree reconstructions estimated direct transmission from an observed source for 42% of cases and provided empirical evidence for multiple co-circulating genetic lineages. Direct transmissions rarely crossed major rivers and cases appeared to align with major road networks, indicating the influence of landscape heterogeneity on rabies dissemination. In addition, comparison to trees informed only by spatio-temporal data demonstrated the necessity of genetic information to correctly infer transmission events. Although based on small number (5 direct comparisons) the data show 56% of source assignments by spatio-temporal inference alone were strongly at odds with the genetic data. The preliminary results shown here demonstrate the potential for transmission tree reconstructions to inform rabies control programmes and provide a model framework for future developments.

5.2 Introduction

The means by which an infectious disease transmits from one host to another is the fundamental process underlying infectious disease dynamics. Understanding the dynamic processes that structure such events is critical to predicting spatiotemporal patterns of incidence and to the design of optimal control strategies (Keeling *et al.*, 2003; Kiskowski & Chowell, 2015; Rees *et al.*, 2013; Tuite *et al.*, 2011). However, since transmission is rarely observed, insights rely mainly on statistical methods to reconstruct transmission histories from available data. Ideally, reducing uncertainty and maximising the accuracy of this reconstruction requires the incorporation of all available sources of data, e.g. exploiting information from epidemiological and genetic perspectives (Cottam *et al.*, 2008). However integration of different types of data

into statistical inferential frameworks is conceptually and technically challenging. In the era of "big-data" (Kao *et al.*, 2014), there is an increasing need for such integrative methods to harness the wealth of information available, in particular the incorporation of highly resolved genetic data obtained from modern sequencing technologies.

Two main approaches relying on powerful Bayesian inference schemes have emerged in the past decade to combine spatial, temporal and pathogen genetic data. The first relies on phylodynamic methods, as explored in Chapters 3 & 4, which use coalescent models to simultaneously measure epidemiological processes and pathogen evolution. Although this approach is robust to sampling intensity, inference is constrained by the use of simple epidemiological models that don't capture more complex, stochastic population effects and cannot easily be related to real epidemiological processes (Kao *et al.*, 2014; Rasmussen *et al.*, 2011). Recent implementations have incorporated more advanced models such as the birth-death model (Stadler, 2009) or Susceptible-Infected-Recovered (SIR) population model (Rasmussen *et al.*, 2011) but still cannot account for more specific heterogeneities in susceptibility or contact patterns (Kao *et al.*, 2014).

In contrast transmission tree reconstructions take advantage of explicit models of transmission to account for host population structure and underlying epidemiological processes, achieving the best approximation of "who infected whom" strengthened by inferences from genetic data (Jombart *et al.*, 2014; Mollentze *et al.*, 2014b; Morelli *et al.*, 2012; Ypma *et al.*, 2012, 2013). This discriminatory power offers the potential to identify how landscape and population processes affect transmission on scales that may not be detectable using traditional phylogenetic approaches. Exploring transmission at this refined scale facilitates efficient estimates of transmission parameters to inform disease control and surveillance. This includes the best estimation of who infected whom (Morelli *et al.*, 2012), the rate of mutation per transmission event (Cottam *et al.*, 2008), the proportion of unobserved cases (Mollentze *et al.*, 2014b), the effective reproductive rate (Jombart *et al.*, 2014) and most likely transmission pathways (Jombart *et al.*, 2014; Mollentze *et al.*, 2014b; Morelli *et al.*, 2012; Ypma *et al.*, 2012, 2013). Moreover, it may provide a means to assess the contribution of introductions to the persistence of acute fatal diseases such as rabies via quantification of their frequency and successful establishment.

The development of transmission tree inference methods is at an early stage and has mostly been applied to epidemic rather endemic situations. However, Mollentze *et al.* (2014b) recently introduced a framework adapted from Morelli *et al.* (2012) to address complexities specific to endemic systems. This included accommodating for the possibility of multiple exogenous introductions (i.e. outside the sampled area), rather than considering an outbreak resulting from a single introduction, and enabling the reconstruction of transmission trees when a proportion of cases are unobserved. Soubeyrand (2014) has further adapted this approach, proposing an alternative approximate MCMC algorithm to improve transmission tree inference and computation times. I applied the approach presented in Soubeyrand (2014) to infer

transmission dynamics in a local (i.e. administrative district scale) endemic rabies system in Tanzania and extended it by incorporating whole genome sequences, contact tracing data and known epidemiological parameters for my study system.

Rabies virus is characterised by a high mutation rate, which means that epidemiological and population genetic processes occur on a similar timescale (Drummond *et al.*, 2005). As such rabies virus provides an insightful system for the elucidation of ecological and evolutionary dynamics. In addition, the memorable nature of rabies transmission events, i.e. bites from rabid animals and often distinctive clinical signs, make rabies amenable to contact tracing (Hampson *et al.*, 2009) and the data for this Chapter contains a proportion of individuals with traced sources of transmission. These properties make rabies an ideal system to study processes underlying transmission dynamics (Biek *et al.*, 2007; Brunker *et al.*, 2012; Hampson *et al.*, 2009) and the level of epidemiological and genetic data available for the study system provides a unique opportunity to take advantage of cutting-edge inference techniques described above.

In the Serengeti District mass dog vaccination campaigns have been delivered annually since 2002 but have failed to achieve a lasting hold on reductions in incidence and eliminate the disease. Transmission tree reconstructions can potentially provide important insights into patterns of local endemic transmission that can better inform control efforts. Whole genome data provides additional information to delineate between cases close in time and space but not necessarily from the same transmission chain, which I demonstrate here by comparing inference from spatiotemporal data to the integrated space-time-genetic approach. In particular, as two major phylogenetic lineages have previously been identified in the Serengeti District (Brunker *et al.*, 2015) I determine the extent to which they co-circulate and share a spatial distribution.

5.3 Materials and Methods

5.3.1 Data

A proportion of the cases used in Chapter 4 was sub-sampled to provide a dataset amenable for computation. This sub-sample consisted of 257 observed cases collected between 1st July 2012 and 31st August 2013, of which 16% had genetic information and 37% had observed contacts (see Fig.5.1). Details for the sequenced cases (n=41) can be seen in Appendix C.1 (samples used in this chapter have an asterisk).

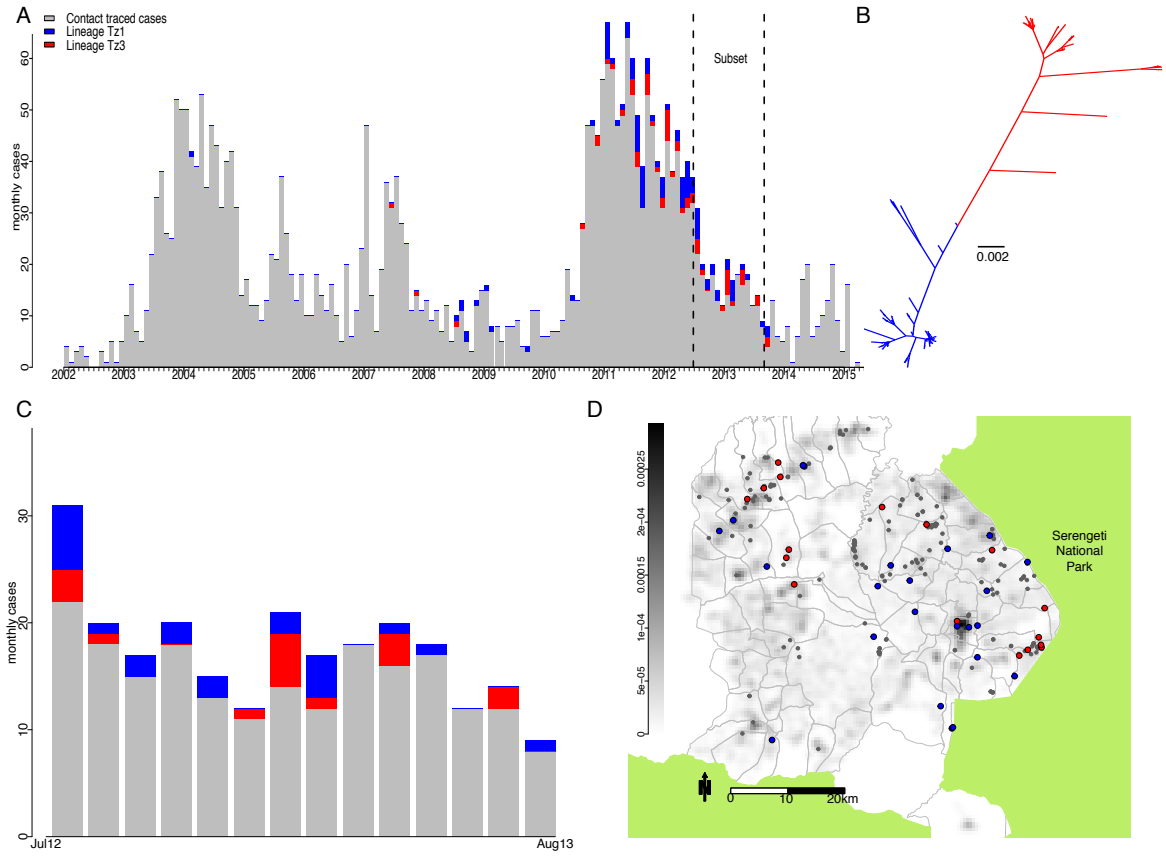


Figure 5.1: Rabies cases recorded in the Serengeti District and subset used for transmission tree reconstructions: A) monthly rabies cases recorded from 2002 to 2015 with 152 whole genome sequenced samples from major phylogenetic lineages Tz1 (blue) and Tz3 (red), unsampled in grey, and window used for transmission trees highlighted; B) Maximum likelihood phylogeny of the 152 genetically sequenced samples indicated in (A); C) monthly rabies cases for subset used in computations; D) spatial distribution of subset cases in the Serengeti District with underlying dog density distribution.

5.3.2 Whole genome sequences

Whole genome sequences (WGS) were generated under an established pipeline described in Chapter 3. A SNP filtering pipeline was implemented as described in Chapter 3. However, in order to incorporate the highest possible level of non-ambiguous genetic information in the model a relaxed consensus calling protocol was adopted. A consensus was formed allowing any base call with a depth of at least 2 and a majority of 51% (50-50 calls were assigned an IUPAC ambiguity code). Any base position with a coverage less than 2 (including gaps in coverage) was assigned the population consensus call for that position. Base positions with ambiguities are automatically stripped from all samples before incorporation into the model, by implementing this relaxed consensus protocol I maximised the genetic information available to inform transmission tree reconstruction.

5.3.3 Transmission tree reconstruction

I applied the approach presented in Soubeyrand (2014) which is based on a genetic-space-time model, combining (i) an individual-based, spatial, semi-Markov SEIR (susceptible, exposed, infectious, removed) model describing the spatio-temporal dynamics of the pathogen, and (ii) a Markovian evolutionary model to define the temporal evolution of pathogen genetic sequences. The resulting model is a state-space model including latent vectors of high dimension (e.g. the transmission tree, infection times, unobserved sequences of the pathogen at the time of transmission). Extensions were made to incorporate contact tracing data and a zero-inflated dispersal kernel. Additionally I made a distinction between the time at which the host is observed as infected and the end time when the host is removed, as per Morelli *et al.* (2012). In the following, I do not detail the whole approach but provide a basic overview of the model and the extensions mentioned above. A more formal description of the extensions can be found in Appendix D.

5.3.4 Model overview

The objective of the model is to infer a transmission tree J that states who infected whom. An observed individual i can become infected at time T^{inf} by a source j that is either another observed individual or an exogenous source (refer to Fig. 5.2 for a diagrammatic presentation). The model calculates the probability that any host j infected another host i from a joint posterior distribution that determines the likelihood of various parameters defining the infection potential of j. These include likelihoods to account for host population processes (SEIR), observed contact information, spatial dissemination and genetic evolution of cases over time. The model structure is also shown in Fig. 5.2. An exogenous source of infection was represented by a unique sequence S_{exo} defined *a priori*, which was designed to be equidistant from all observed sequences. This was constructed as follows:
- if a site is not variable the corresponding nucleotide is used;
- if a site is variable and there is a nucleotide that never appears at the site, this nucleotide is used;
- if a site is variable and all nucleotides are represented at the site one is selected uniformly randomly.

The use of an exogenous sequence and its date allowed me to easily handle unobserved cases, i.e. the missing infecting hosts. In contrast to Mollentze *et al.* (2014b) who used the reconstructed sequence of the most recent common ancestor (MRCA) as S_{exo}, I chose to use a central sequence due to the observation of at least two very divergent genetic lineages circulating in the Serengeti (Chapters 3 & 4. (Mollentze *et al.*, 2014b) approach led to an MRCA sequence genetically biased towards one lineage and affected the estimation of transmission events, artefacts that were overcome using a central sequence. Observed data were the observation times T^{obs}, removal times T^{end}, central location of observed individuals X, ability

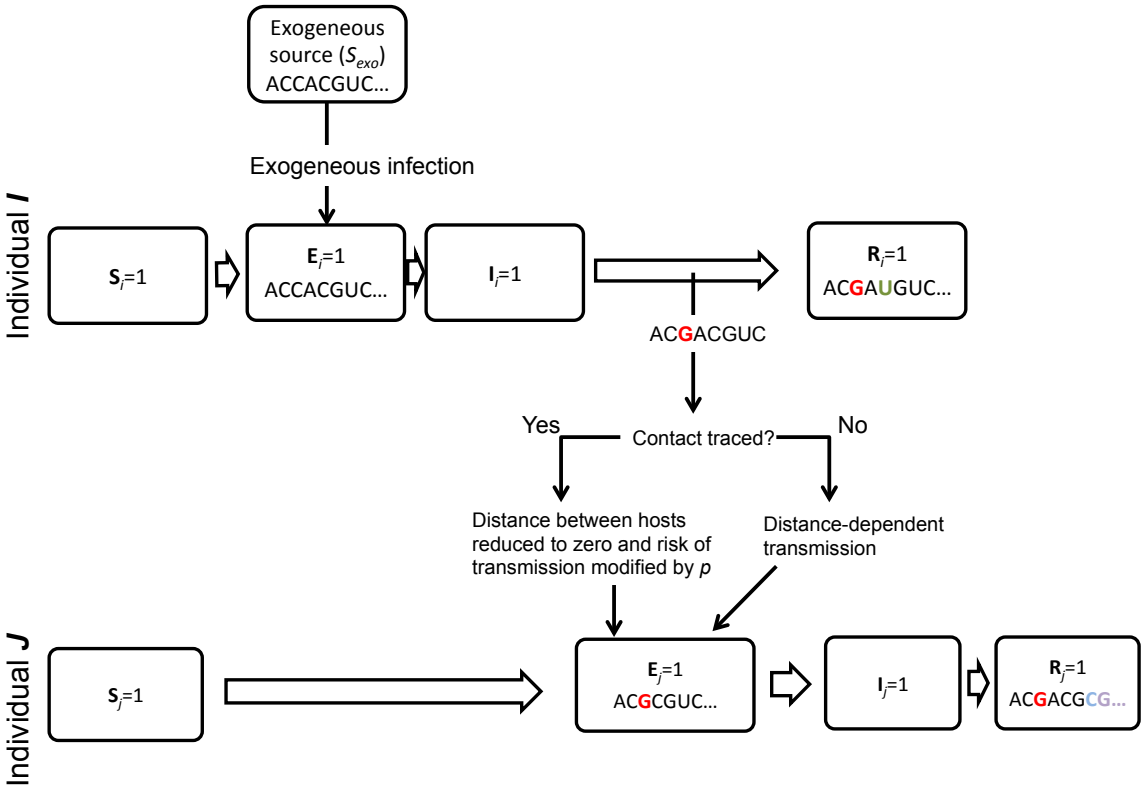


Figure 5.2: Model schematic illustrating the genetic-space-time model combining a semi-Markov SEIR model and a Markovian evolutionary model. Here Individual i is infected by an exogenous source represented by a central sequence (S_{exo}) "ACCACGUC...". Individual i becomes infectious and infects j at a point in time when the sequence in i has evolved (see C at the 3rd base had mutated to G). The probability of transmission *J(i)* is informed by contact tracing information if observed, which reduces the dispersal distance to zero and alters the probability by a fixed value of p. If not contact tracing is observed transmission is distance-dependent. After transmission both sequences in i and j continue to evolve independently. Modified with permission from (Soubeyrand, 2014).

to spread the disease i.e. a spreader (domestic dog, cat or wild carnivore) or a dead-end host (livestock), observed sequences collected from individuals at the time of death S^{obs}, the sequence of the disease reservoir S_{exo} and contact tracing information. In addition, previous studies provide informative prior information for epidemiological parameters including the duration of incubation and infectious periods and the spatial dispersal kernel (Hampson *et al.*, 2009).

5.3.5 Posterior distribution

I consider the joint posterior distribution $p(J, T^{inf}, L, D, \theta \mid data)$ of the transmission tree J, infection times $T^{inf} = (T_1^{inf}, \ldots, T_n^{inf})$, exposed (or latency) durations $L = (L_1, \ldots, L_n)$,

infectious durations $D = (D_1, \ldots, D_n)$ before observations, and parameters θ that contains infection and dispersal parameters $\alpha = (\alpha_0, \alpha_1, \alpha_2) = (\alpha_0, \alpha_1, (\alpha_{2,1}, \alpha_{2,2}, \alpha_{2,3}))$, latency parameters $\beta = (\beta_1, \beta_2)$, infectiousness parameters $\delta = (\delta_1, \delta_2)$, mutation parameters $\mu = (\mu_1, \mu_2, \mu_3)$ and the date t_{exo} of the exogenous sequence S_{exo}. The full equation for the posterior distribution is shown in Appendix D and includes the following components:

i) Genetic likelihood; $p(S^{obs} \mid J, T^{inf}, L, D, \theta, T^{end}, X, S_{\text{exo}})$

Pathogen sequences were considered to evolve with time under a 3-parameter Kimura substitution model (Kimura, 1981). Mutations can only occur between four possible nucleobases (ACTG) and positions with an ambiguous base are stripped from the observed sequence alignment. I refer to Mollentze *et al.* (2014b) for the expression of the genetic likelihood and its approximation.

ii) Contact tracing

Contact tracing was introduced into the model by incorporating a time-varying dispersal kernel $\tilde{w}$ such that when i and j are in contact the distance between them is reduced to zero and a multiplicative factor ρ, specified *a priori*, modifies the risk of transmission. See Appendix D for further details.

iii) Transmission likelihood; $p(J, T^{inf} \mid L, D, \theta, T^{end}, X, S_{\text{exo}})$

Each host has the same chance (1/I) to be the first infected by an exogenous source and its infection time is assumed to be less than or equal to the first observation time. The probability of subsequent infections is drawn from a mixture model that gives the probability density that the infection arose from each type of source (exogenous, direct or direct contact-traced). This is represented by the following equation (where 1 is an indicator value):

$$
\begin{aligned}
& p\left(J(i), T_i^{inf} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{inf}, L, D, \theta, T^{end}, X, \mathcal{C}\right) \\
&= p\left(J(i), T_i^{inf} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{inf}, L, \theta, T^{end}, X, \mathcal{C}_i\right) \\
&= \exp\left(-\alpha_0 (T_i^{inf} - T_1^{inf}) - \int_{T_1^{inf}}^{T_i^{inf}} \sum_{\substack{j=1 \\ j \neq i}}^{I} \alpha_1 \mathbf{1}(T_j^{inf} + L_j \leq t \leq T_j^{end}) A_j w(x_j - x_i) dt \right. \\
&\qquad \left. - \sum_{\substack{j \in \mathcal{C}_i \\ j \neq J(i)}} \epsilon \rho \alpha_1 \mathbf{1}(T_j^{end} \leq T_i^{inf}) A_j w(0) \right) \\
&\quad \times \Big(\alpha_0 \mathbf{1}\{J(i) = 0\} \\
&\qquad + \alpha_1 \mathbf{1}(T_{J(i)}^{inf} + L_{J(i)} \leq T_i^{inf} \leq T_{J(i)}^{end}) A_j w(x_{J(i)} - x_i) \mathbf{1}\{J(i) \neq 0 \text{ and } J(i) \notin \mathcal{C}_i\} \\
&\qquad + \rho \alpha_1 \mathbf{1}(T_{J(i)}^{inf} + L_{J(i)} \leq T_i^{inf} \leq T_{J(i)}^{end}) A_j w(0) \mathbf{1}\{J(i) \neq 0 \text{ and } J(i) \in \mathcal{C}_i\} \Big)
\end{aligned}
\tag{5.1}
$$

The exponential term is the probability that host i has not been infected between times T_1^{inf} and T_i^{inf}, and the second term is the probability density that host i has been infected by $J(i)$ at time T_i^{inf}. Here, if $J(i) > 0$ the source is observed, while the source is external to

the dataset (an exogenous source) if $J(i) = 0$. α_0 is the infection strength of the exogenous sources, assumed to be constant in time and space, α_1 is the infection strength of an observed source, and w is a parametric dispersal kernel. This kernel is assumed to be a zero-inflated power-exponential kernel parametrised by $\alpha_2 = (\alpha_{2,1}, \alpha_{2,2}, \alpha_{2,3}) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \times [0,1]$ and satisfying, for all $x \in \mathbb{R}^2$:

$$w(x) = \alpha_{2,3} + (1 - \alpha_{2,3}) \frac{\alpha_{2,2}}{2\pi(\alpha_{2,1})^2 \Gamma\left(\frac{2}{\alpha_{2,2}}\right)} \exp\left\{ - \left(\frac{||x||}{\alpha_{2,1}} \right)^{\alpha_{2,2}} \right\}. \tag{5.2}$$

iv) Distribution of latency and infectious durations; $p(L, D \mid \theta, T^{end}, X, S_{\text{exo}})$
As per Mollentze *et al.* (2014b) the distribution of latency durations and infectious durations were modelled according to independent gamma distributions parameterised by contact tracing data from the Serengeti District (Hampson *et al.*, 2009).
v) Prior distribution of parameters independent of the explanatory variables; $p(\alpha_0, \alpha_1, \beta, \delta, \mu, t_{exo})p(\alpha_2)$
Independent prior distributions were used for all parameters, see Table 5.1 for details.

Table 5.1: Prior distributions and other model specifications

	Distribution	Prior
Incubation period	$L_i \underset{\text{indep.}}{\sim} \Gamma(\beta_1, \beta_2)$	(β_1, β_2) fixed such that $E(L_i) = 22.1$days $sd(L_i) = 21.2$days
Infectious period	$D_i = T_i^{obs} - T_i^{infectious}$ $\underset{\text{indep.}}{\sim} \Gamma(\delta_1, \delta_2)$ $T_i^{end} - T_i^{obs} = 2$ days	(δ_1, δ_2) fixed such that $E(D_i) \approx 1.5$days $sd(D_i) \approx 0.5$days
Dispersal	$x_i - x_j \underset{\text{indep.}}{\sim} w = w(\cdot; \alpha_2)$: (zero-infl. exp.-power kernel)	Prior over the mean dispersal distance $\bar{d}(\alpha_2) = (1 - \alpha_{2,3})\alpha_{2,1}\frac{\Gamma(3/\alpha_{2,2})}{\Gamma(2/\alpha_{2,2})}$ $\bar{d}(\alpha_2) \sim \Gamma(d_1, d_2)$ such that $E(\bar{d}(\alpha_2)) = 0.88$km $sd(\bar{d}(\alpha_2)) = 0.10$km
Strength exo. sources	α_0	$\alpha_2 \sim$ Exponential(rate $= 1/100$)
Strength obs. sources	α_1	$\alpha_1 \sim$ Exponential(rate $= 1/100$)
Contact	(ρ, ϵ)	$(\rho, \epsilon) = (100, 0.1\text{day})$
Transition	μ_1	$\mu_i \underset{\text{indep.}}{\sim}$ Exponential(rate $= 1/(2 \times 10^{-6})$)
Transversion of type 1	μ_2	such that $E(\mu_i) = 2 \times 10^{-6}$)
Transversion of type 2	μ_3	per day per nucleotide
Exogeneous sequence	A sequence S_{exo} built such that it is at equal distance from all observed sequences	
Time of the exo. sequ.	t_{exo}	$t_{\text{exo}} \sim$ Normal$(mean = 0, sd = 15000)$

5.3.6 MCMC chains

20 parallel MCMC chains were run, with each chain subsampled every 200 iterations. A burn-in of 20,000 iterations was applied and chains run until convergence was achieved.

5.3.7 Landscape features

Spatial data for major roads, rivers and dog density were manipulated and mapped in R version 3.2.0 (R Core Team, 2015). Simple measures were applied to assess the effect of landscape features on transmission events. Intersections between rivers and the straight line connections between infected hosts and their observed source were measured using the rgeos package (Bivand & Rundel, 2014) in R. I calculated the proximity of infected cases to a major road using the gDistance package (van Etten, 2015).

5.4 Results

5.4.1 Transmission tree inference from space-time-genetic data

A posterior sample of size of 641 was obtained, which was adequate to achieve convergence on all parameters (posterior distributions can be seen in Appendix D). The model reconstructed transmissions for a total of 257 cases between 1st July 2012 and the 31st August 2013, of which 42.42% were predicted to have direct links to observed sources and the remainder had sources exogenous to the observed dataset. The mean incubation period for directly inferred transmissions using the space-time-genetic model was 21.65 days and the mean infectious period was 3.44 days (Fig.5.3).

Maximum a posteriori probability (MAP) estimates for inferred sources of cases (both direct and exogenous) had a median value of 0.98 (2.5^{th} and 97.5^{th} quantiles: 0.91 and 1) and 97% had a MAP $>$0.5. MAP transmissions for each infected host are shown through time in Fig.5.4. The two major lineages Tz1 and Tz3, shared a spatial distribution across the district, co-occurring in both time and space. Direct reconstructions were estimated for 15/41 sequenced cases, with 6 estimated to have a genetically sequenced progenitor, which were all from the same lineage as the infected host.

Cases with estimated MAP transmissions less than 0.5 (7/257) had posterior distributions split across competing observed sources. In 5/7 cases this prevented the model from selecting a single observed source as the most probable progenitor, instead assigning an exogenous source despite a combined overall probability that the infection came from within the observed

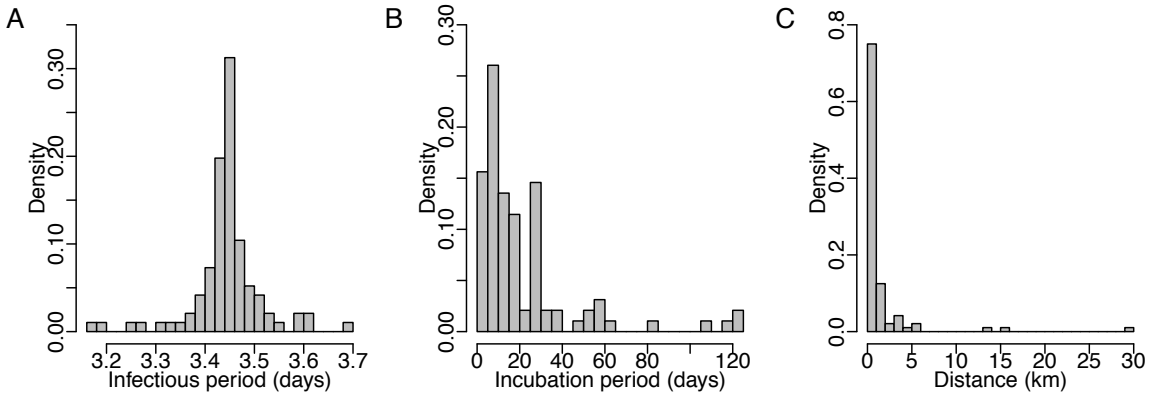


Figure 5.3: Posterior distribution of A) incubation periods; B) Infectious periods and C) transmission distances between cases that were inferred to be directly connected using the space-time-genetic model with the highest posterior probability

data (see Fig.5.5). In these instances the infected host and potential sources were within narrow, densely sampled spatio-temporal windows and did not have observed sequence data. Most infection events occurred between hosts in the same location and the median distance excluding zero distance events was 0.94km (Fig.5.3). It was rare for direct transmissions to traverse rivers, I found only 2 instances (out of 96 direct transmissions) when a straight line connection crossed a river. Infected cases appear to align to road networks: 65% were less than 5km from a major road and 36% were within 1km. This was true for cases with direct or exogenous sources (Fig.5.6). Cases with exogenous sources were distributed across the region and the proportion of exogenous cases did not show an overall decrease through time (Fig.5.7).

5.4.2 Comparison to inference without genetic data

I compared results from tree reconstructions to those obtained from reconstructed epidemic trees based on the spatiotemporal proximity of cases (Hampson *et al.*, 2009). In contrast to my model which allows for the possibility of exogenous sources, Hampson *et al.* (2009) forced cases to find connections to observed sources and therefore the most likely source may still be highly unlikely. There were 9 transmission events estimated by the spatiotemporal reconstruction for which genetic data was available for both host and source. In 56% (5/9) of these cases, the progenitor assigned using spatiotemporal information was from a genetically divergent lineage to the infected host, see Table 5.2. The log likelihood scores for these transmission events were relatively low (see distribution of log-likelihoods in Appendix) but it is difficult to identify a cutoff for very likely or unlikely progenitors. In contrast, the genetically informed reconstructions assigned an exogenous source to 4 of these cases with high probability ($>$0.7) and assigned 1 to an alternative sequenced progenitor from the same lineage.

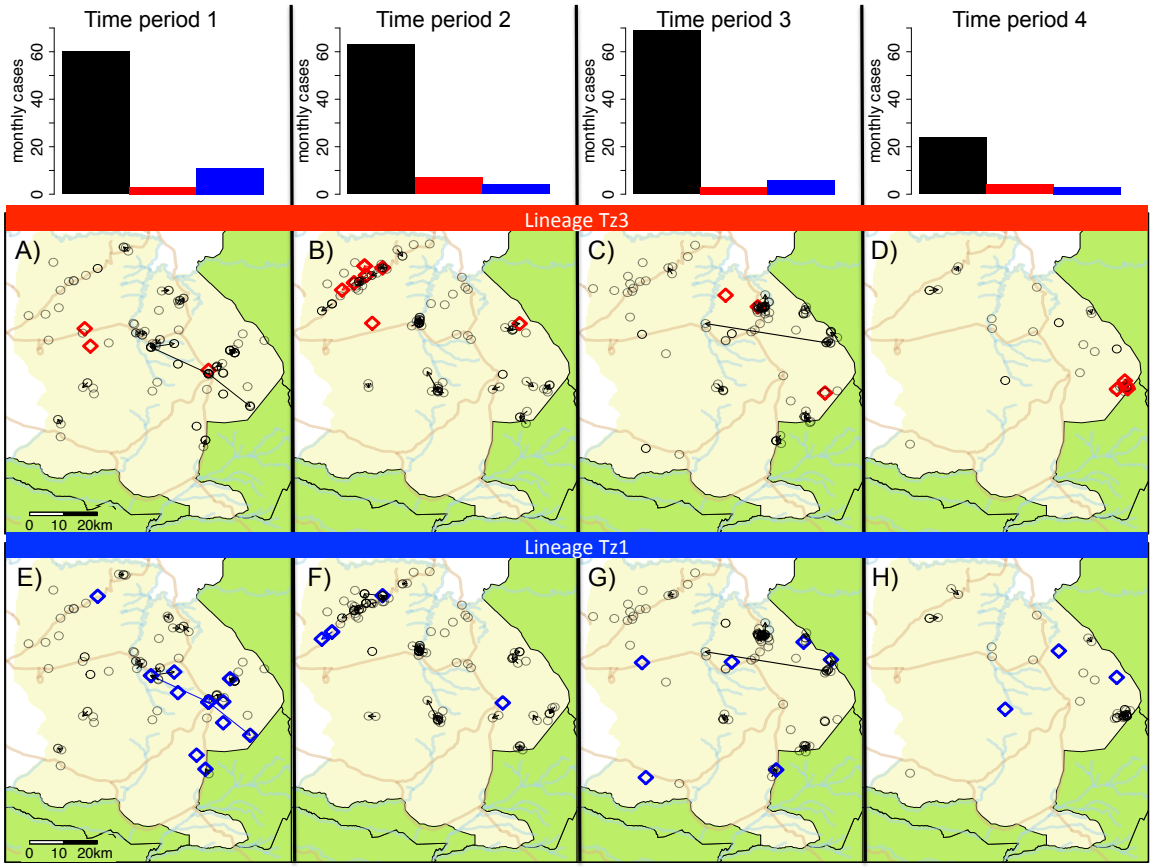


Figure 5.4: Most probable transmission events in each quarter of the sampled period in the Serengeti District, shown with major roads (brown lines) and rivers (light blue lines). The first row of maps (A-D) highlights observed cases with sequence data belonging to lineage Tz3 in red, while the second row (E-H) shows lineage Tz1 cases in blue. Black circles are observed cases without sequence data. Note: in both rows the same cases are plotted i.e. all cases within the time period but each lineage is highlighted on a separate row to aid with visualisation. Arrows, weighted by the strength of posterior support, indicate direct transmissions between observed hosts and are coloured if the source of infection had sampled genetic data. Symbols not proceeded by an arrow are cases where the most likely progenitor was an exogenous source. The number of cases in each quarter for unobserved (black), Tz1 (blue) and Tz3 (red) cases is shown at the top. (A small amount of jitter has been added to points less than 300m apart.)

5.5 Discussion

In this chapter I have demonstrated the utility of a powerful Bayesian inference scheme to estimate transmission trees in an endemic system using combined genetic and epidemiological data. Results highlight the necessity of genetic data to correctly infer transmission events in a local (district-level) endemic system sampled within a limited spatio-temporal window. The inherent complexity in endemic systems generates considerable challenges to uncovering underlying transmission dynamics. However, by amalgamating all available data into a single inference framework I was able to expose relationships between samples occurring very close

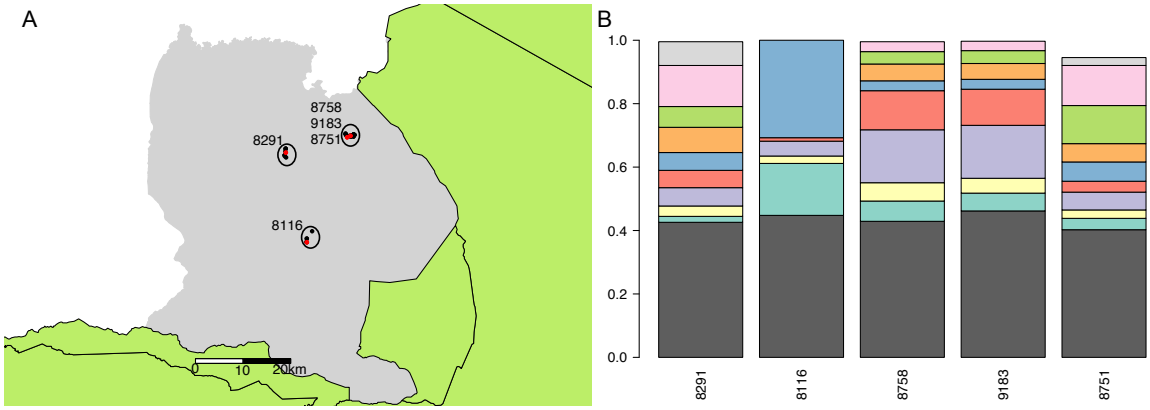


Figure 5.5: Rabies cases with many possible observed sources: A) infected hosts shown in red with possible sources in black, each cluster of cases is labelled with the infected host ID; B) posterior distributions for each case with probabilities shown for the top 10 estimated sources, including an exogenous source in dark grey and possible observed sources in other colours. The overall probability of an observed source is greater than the probability of an exogenous source but no single observed source had a majority probability and therefore each host was assigned an exogenous source.

Table 5.2: Transmission tree reconstructions for hosts with genetic information using 1) spatiotemporal proximity (i.e. genetic information not used) between cases to assign progenitors in a maximum likelihood approach (Hampson *et al.*, 2009) and 2) space-time-genetic inference to assign most probable progenitors in an integrated bayesian inference scheme. Samples are labelled according to the phylogenetic lineage they belong to and results from each algorithm are shown: log likelihood results from the maximum likelihood tree using spatiotemporal reconstruction, and posterior probabilities from bayesian space-time-genetic reconstructions.

		Space-time			**Space-time-genetic**		
Infected host	**Infected lineage**	Progenitor	Lineage	Loglik	Progenitor	Lineage	Post prob
RV3135	b	RV3146	a	-17.73	RV3131	b	0.57
RV3140	a	RV3124	b	-14.11	non-observed	—	0.99
RV3057	a	RV3052	b	-19.11	non-observed	—	0.99
RV3087	a	RV3085	b	-20.12	non-observed	—	0.9
RV3086	a	RV3085	b	-20.22	non-observed	—	0.73

together in space and time.

Tree reconstructions highlighted the capacity of genetic data to infer transmission events, distinguishing between lineages that were circulating in the same locations. While the presence

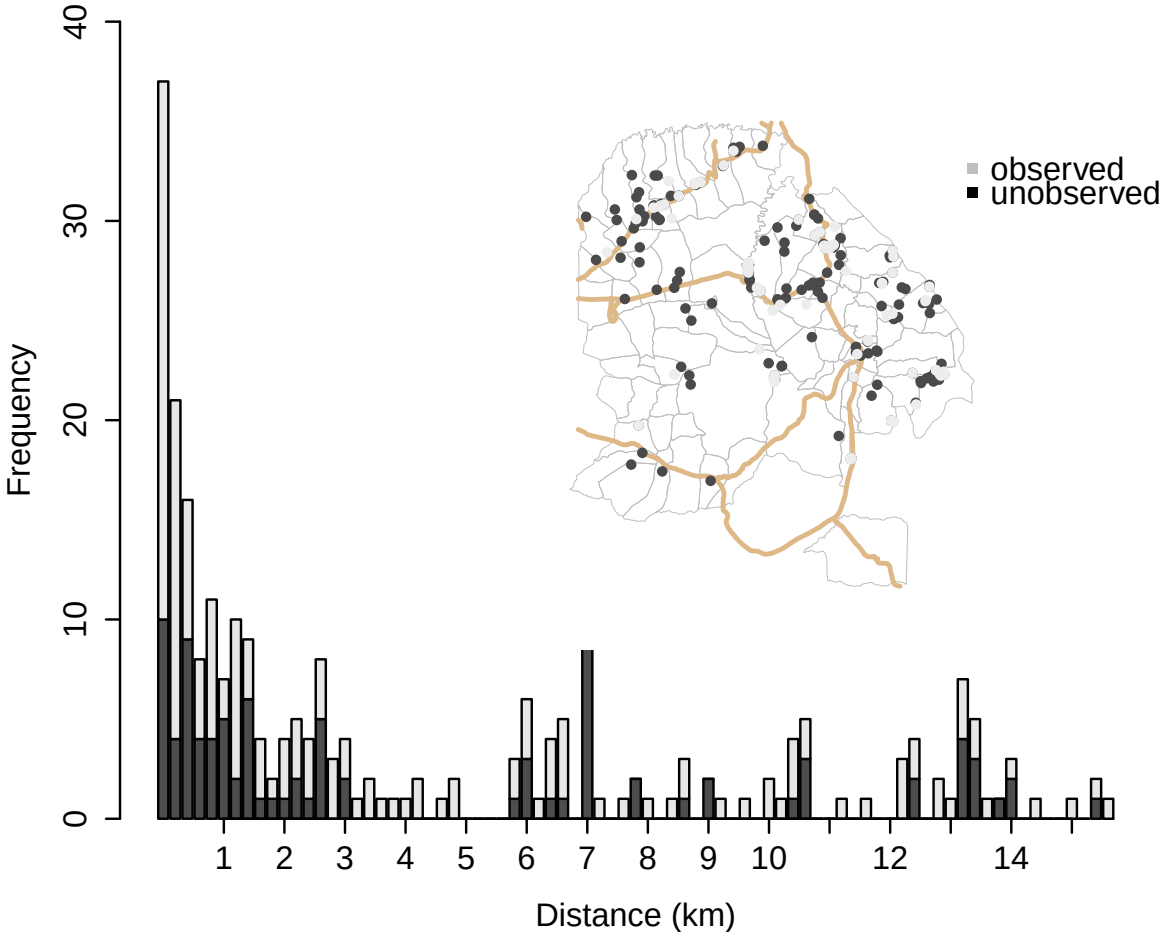


Figure 5.6: Proximity of infected hosts with observed (light grey) and unobserved (dark grey)sources to major roads shown in a stacked histogram.

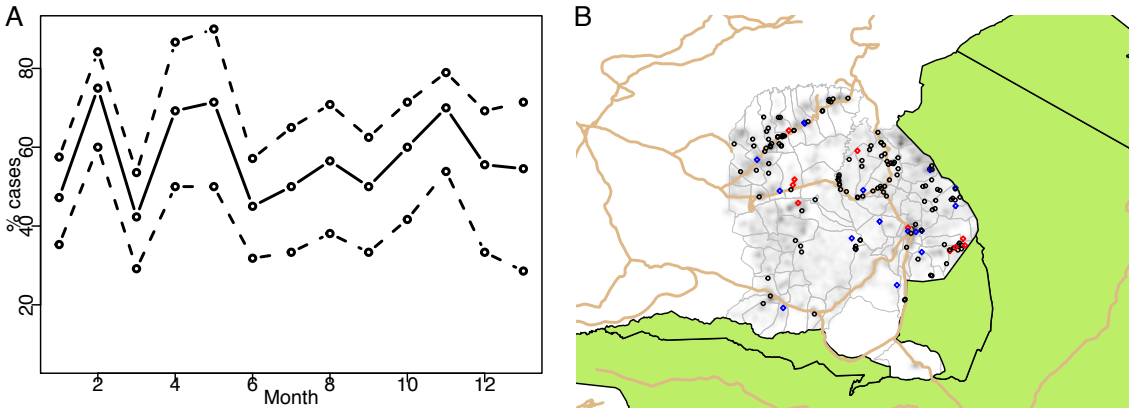


Figure 5.7: A) Proportion of cases assigned an exogenous source through time and B) their spatial distribution overlaying dog density and roads.

of major lineages can also be identified using phylogenetic analyses (Chapters 3 & 4), at fine spatio-temporal resolutions the transmission tree can begin to decouple from the phylogenetic

tree (Kao *et al.*, 2014). This makes the inference of who infected whom less reliable using phylogenies as the timing of nodes represents coalescent events, not times of transmission (Ypma *et al.*, 2013). As such the ability of transmission tree reconstructions to integrate additional epidemiological data and contact structure constitutes a major advantage to identify patterns of direct transmission that are hard to discern from phylogenetic analysis. The presence of two very distinct lineages in the Serengeti District (or more depending on the level of subdivision) appears to be a result of endemic circulation over a number of decades (Chapter 3). Using transmission reconstructions, I show that these two lineages consistently co-circulate in the same population.

RABV lineages in other host species (e.g. raccoons, skunks, brown bats) commonly exhibit distinct geographical structuring (Biek *et al.*, 2007; Bourhy *et al.*, 2008; Szanto *et al.*, 2011; Torres *et al.*, 2014) but a few studies have identified sympatric areas (Barton *et al.*, 2010) or certain lineages with wide-spread distributions (Bourhy *et al.*, 2008; Nadin-davis *et al.*, 2010). In Barton *et al.* (2010) two sympatric skunk lineages occupying several counties in the central Great Plains of North America were found to have distinct viral properties resulting in transmission patterns characteristic of epidemic (South Central Skunk rabies strain) and endemic (North Central Skunk rabies strain) transmission patterns. These differences may be a result of different host histories (one lineage evolved from bats and the other from dogs). In Chapter 3 I found multiple lineages in regional areas in Tanzania, a pattern attributed to human mediated movements. Given the ties between dog and human ecology It is likely that humans have a significant influence on the mixture and persistence of dog rabies lineages in the Serengeti District.

It is currently unclear whether the two Serengeti lineages persist as a result of repeated introductions into the district from the larger endemic area which they occupy, or whether they have both continued to circulate even at low levels within the district. Understanding the transmission dynamic that have lead to their persistence is important for effective control. For example, persistence as a result of repeat introductions from outside the district may require improved surveillance and interventions to restrict influx from common sources. Historical genetic material from the Serengeti District indicates that the lineages were present 20 years ago (Chapter 3) and throughout the more recently sampled period (Fig.5.1). This suggests that neither drift nor selection have allowed one lineage to outcompete the other- likely mediated by the high availability of susceptible hosts. Once control measures have reduced the susceptible population to some critical threshold it is likely that local extinction of one or both of these lineages will occur. Genetic surveillance can play a crucial role in tracking the progress of control efforts and identifying problematic areas or sources of incursion in the later stages of elimination.

The model currently assigns sources as exogenous when transmissions involve unobserved ancestors, which includes both indirect (i.e. when there are missing intermediate hosts) and

true exogenous transmissions. The proportion of exogenous sources estimated did not show an overall decrease through time, hinting that there may be a consistent flow of infections from outside sources. However, a distinction between true exogenous sources and indirect sources is needed to clarify this. Mollentze *et al.* (2014b) used a post-processing algorithm to delineate between each type, which was computationally intensive and problematic when sampling intensities were low. I attempted to directly infer direct, indirect and exogenous transmissions but was unable to fine-tune at this stage. However, in future it should be possible to make this distinction and estimate the contribution of true exogenous sources to the system dynamic.

Without knowledge of the underlying genetic data it is unlikely that spatiotemporal data alone can correctly infer transmission trees in the Serengeti District. Comparisons between reconstructions using different methods indicated that 56% of cases for which genetic information was available (5 out of 9) were assigned a source in strong disagreement with their sequence data. The majority of MAP estimates were able to clearly designate a source, whether direct or exogenous, with high probability but in 3% of cases the space-time-genetic algorithm was unable to assign an observed source from several that were feasible. In such cases the model assigned an exogenous source despite an overall majority probability that the source came from within the observed data (shown in Fig.5.4). These cases had no observed sequence data, which would have provided additional resolution to identify the most likely progenitor, in particular eliminating highly unlikely progenitors from different lineages. While reconstructions based exclusively on spatio-temporal data may be suitable for epidemic scenarios it is apparent that genetic information is necessary to infer transmission events accurately in this endemic system. It is unclear how generalisable this is to other endemic pathogens, but the presence of co-circulating dengue 1 virus lineages has also been observed in large metropolitan areas (Ospina *et al.*, 2010; Raghwani *et al.*, 2011) suggesting it may be a characteristic of some endemic systems.

Landscape heterogeneity may influence transmission events and could provide informative prior information to increase the accuracy of transmission tree reconstructions. A future goal of this model is to incorporate heterogeneity into the inference of transmission events. As the main results from Chapter 4 support rivers and roads as influences on RABV dispersal I performed some basic exploratory analysis to compare the presence of these features with transmission events inferred here. Direct transmissions across rivers were only observed in 2% of cases, hinting that they limit RABV dispersal to some extent. However, null models are needed to test any significance statistically and assess the permeability of rivers to diffusion over a longer time period before any decisive measures can be taken regarding control. A more thorough quantification of their effects could be achieved through landscape permutations and simulation studies, potentially providing prior information for tree reconstructions e.g. adjusting probabilities of transmission across a river by x-fold. I also measured the proximity of infected individuals to roads and found that 65% of all infected cases occurred within 5km

of a road and 36% were within 1km, suggesting that transmissions are influenced by human connectivity. Major roads are most likely to be in areas with high human population density, which correlates with dog density (see Fig.5.1D) and may reflect the availability of susceptible hosts. In addition to rivers and roads, there are many landscape features and processes that may structure transmission events. Transmission tree reconstruction may help to tease apart some of these influences and, in turn, may benefit from their incorporation in probability estimates.

This analysis was used as a preliminary test of the model to reconstruct transmission in a local endemic system. As such there are improvements to future models that could be made based on the results. Already mentioned are possible developments to infer indirect transmissions between observed sources and include landscape heterogeneity. To represent unobserved/ exogenous sources of infection I generated a central sequence to relate any observed sequence equally to the reservoir. This was to overcome the issue of finding a reconstructed MRCA to represent a diverse group of sequences without bias towards certain clades. Alternative approaches would be to use multiple common ancestors corresponding to different lineages or allow varying substitution rates from the MRCA sequence to each clade. A major advantage of transmission tree algorithms over a coalescent model approach is the ability to incorporate explicit models of transmission. In this Chapter an SEIR model was utilised but alternative model structures could easily be explored within this framework, including the consideration of vaccinated individuals in the population.

Using a subset of data with a proportion of genetically sampled and contact traced cases I was able to reconstruct direct transmissions in 42% of cases. My results show the potential for space-time-genetic inference to uncover transmission dynamics in the Serengeti and provide exciting prospects for future research. I was able to highlight several areas for future development including the inference of indirect transmission events between observed cases, the incorporation of landscape heterogeneity and improvements to prior specifications. Future models should be able to infer longer chains of transmission by allowing for unsampled intermediate cases, offering a more realistic overview of connectivity between cases. This can provide additional information to determine how RABV persists in a local scale, including the frequency of invasions and variation in transmission events as an effect of landscape heterogeneity. Such data can be used to inform improved control and surveillance of rabies virus in areas where circulation has persisted for decades.

CHAPTER 6

General Discussion

The spread and persistence of infectious disease is an inherently spatial process influenced by landscape and population heterogeneities in complex environments. Research based on assumptions of homogeneous mixing no longer holds in the modern paradigm of infectious disease dynamics and efforts have shifted to explicitly incorporate heterogeneities in population and landscape processes, for example using spatially explicit mathematical (Lloyd & May, 1996; Meentemeyer *et al.*, 2011; Panjeti & Real, 2011) or network models (Gardy *et al.*, 2011; Smith *et al.*, 2002). However, spatial variation is difficult to measure and quantify in natural systems (Levin, 1992) and therefore little is known about what spatial processes are important. In Chapter 2 we introduced the re-emerging field of landscape epidemiology, an interdisciplinary approach to disease ecology that takes advantage of a range of analytical tools to study the causes and consequences of spatial variation in host-pathogen systems. The accessibility of whole genome sequencing for pathogen populations, particularly small and rapidly evolving viruses like rabies, has revolutionised the study of evolutionary/ecological interactions, advancing our understanding of disease transmission and potential to interrupt spread and persistence. This thesis directly explored the approaches available to integrate genetic data with landscape ecology research to determine the underlying processes influencing the spread and persistence of rabies, in particular, harnessing the power of whole genome information. In the following sections we synthesise the results from each chapter and discuss the larger concepts, implications and limitations pertinent to landscape epidemiological studies and disease control.

6.0.1 Concept of scale

The spatial scale at which transmission is affected by ecological processes and which is most important for effective control is a question central to epidemiology. Spatial patterns are correlated and change with scale and therefore to understand the impact of landscape structure on disease dynamics one must consider multi-scale information (Wu, 2004). Matching the scale of control with the spatio-temporal scale of disease dynamics is critical to effectively interrupt transmission.

An overarching theme of this thesis has been to explore the degree to which spatial structure can be detected in endemic populations and the consistency of predictors across spatial scales. As cited by Turner (1989), "the measurement of spatial pattern and heterogeneity is dependent upon the scale at which the measurements are made". It is important that several spatial scales are explored to determine the level at which non-linearities in transmission are strongest and what level is most appropriate for effective control or at what scale natural barriers can be exploited. For example, rivers have been implicated as barriers to wildlife rabies dispersal (Ball, 1985; Bingham *et al.*, 1999; Bourhy *et al.*, 1999; Cullingham *et al.*, 2009; Smith *et al.*, 2002) but may become less important at larger scales (Biek *et al.*, 2007). We discovered scale-specific effects of landscape attributes in Chapter 4, where varying levels of discretisation

were applied to represent differing spatial scale. It was apparent from this analysis that landscape heterogeneity had least impact at the smallest spatial scales, predominantly because other (potentially behavioural) heterogeneities operate at these scales and may make any landscape level impacts difficult to detect. There was, surprisingly, detectable structure at small scales (within-district level), for example indicating rivers as barriers to dispersal and roads as facilitators (Chapter 4). In a system where hosts are so intimately attached to human populations small-scale signals are unexpected as humans connectivity often exacerbates the spread of disease (Cleaveland *et al.*, 2007; Greger, 2007; Weiss & McMichael, 2004) and is expected to overcome many natural obstacles to dispersal e,g, bridges/boats across rivers. Our phylodynamic reconstructions did highlight instances of large-scale translocations mediated by humans, suggesting that the impact of human interference on rabies is most important at larger scales.

The question of scale is also relevant to the collection of data since the resolution of data necessary for empirical studies is determined by the scale of interest. As argued in Chapter 3 and shown in Chapter 5, whole genome resolution is required to characterise very local rabies transmission events but partial sequence data may be adequate to uncover larger scale patterns. Therefore, to design genetically informed surveillance programmes, consideration of the resolution required to answer the specific questions of interest is required. This may be very high resolution (WGS) to track local chains of transmission and targeted epidemiological investigations or only partial sequence data may be required if incursions are suspected from other countries or to monitor the elimination of the diverse genetic lineages through time.

6.0.2 Endemic vs epidemic transmission cycles

Infectious diseases, particularly those of zoonotic origin, are often considered in terms of their pandemic potential e.g. SARS, Ebola, H5N1. By comparison endemic zoonoses are under-recognised as a public health problem and thus overlooked in many countries' control priorities (Halliday *et al.*, 2015; Maudlin *et al.*, 2009). This filters down to the research level, in general resulting in a poor understanding of endemic zoonotic pathogen dynamics.

This thesis has highlighted some of the complexities inherent in endemic systems, which make spatial patterns and processes difficult to extract. In particular, the co-circulation of two genetically divergent lineages in the Serengeti complicated transmission tree reconstructions, rendering inference from spatiotemporal incidence less reliable. At the the same time this also made it possible to detect wrongly assigned sources of transmission and demonstrated the value of genetic data. The question still remains as to how rabies, an acute and fatal disease, manages to persist despite consistent evidence supporting a low overall reproductive rate, R_0. Transmission appears to be mainly localised, with mean distances between infections estimated to be less than 1km (Chapter 5) but there were a large proportion of uncharacterised

sources of infection that could reveal more about the dynamic. Crucially our genetic sampling was missing information beyond the Serengeti District, i.e. encompassing the surrounding area. Persistence in the Serengeti may well be a consequence of forces acting at this larger spatial scale, with spread from rabid dogs in neighbouring districts potentially facilitating "rescue effects" that maintain circulation in smaller areas liable to localised extinction (Metcalf *et al.*, 2013).This seems a plausible hypothesis given that past RABV incidence in the Serengeti has been reduced to extremely low levels yet both genetic lineages managed to persist (see results in Chapter 5.

6.0.3 Measuring landscape heterogeneity

Methods to explore landscape heterogeneity are currently a limitation and approaches to generate null hypotheses to test the effects of landscape heterogeneities are needed. Current phylogenetic approaches assume that landscape structure has an observable impact on dispersal patterns and population structure without specifically knowing if this is true. In our phylodynamic approach IBD patterns explained a large proportion of the observed variation in diffusion rates but often landscape resistance distances were marginally better predictors. Attributing residual variation from IBD structure to spatial heterogeneities is challenging and currently lacks statistical power. Research would greatly benefit from simulation studies to explore known models of landscape heterogeneity and their effect on phylodynamic signatures. For example, can genetic discontinuities be identified in data simulated from known dispersal scenarios when there is a pre-defined barrier? Simulations could also be used to determine thresholds of the ability to distinguish landscape heterogeneities from isolation by distance (IBD) patterns.

6.0.4 Reconstructing transmission

The ability to trace transmission pathways of diseases is an increasingly desirable outcome of epidemiological studies, particularly with regards to directly influencing the control of epidemics (Cottam *et al.*, 2008; Haydon *et al.*, 2003; Jombart *et al.*, 2014; Morelli *et al.*, 2012; Ypma *et al.*, 2012, 2013) and understanding dynamics at different scales of transmission (Hughes *et al.*, 2012; Orton *et al.*, 2013). Reconstructing who-infected-whom remains challenging given that most transmission events are unobserved and those that are often lack the spatial-temporal-genetic resolution needed to accurately quantify the event. This includes empirical estimation of mutation rates from directly transmitted cases and the comparison of microevolution at different scales (see 6.0.5).

Two main but not mutually exclusive approaches have emerged in the past decade to combine spatial, temporal and pathogen genetic data. The first relies on phylodynamic methods,

demonstrated in Chapter 4, which use coalescent models to simultaneously measure epidemiological processes and pathogen evolution, allowing estimations of relative effective population size (Drummond *et al.*, 2005), mutation parameters (Drummond *et al.*, 2002), and rate of spatial spread (Pybus *et al.*, 2012). Although this approach is robust to sampling intensity, inference is limited by the use of simple epidemiological models that don't capture more complex, stochastic population effects and can't easily be related to real epidemiological processes (Kao *et al.*, 2014; Rasmussen *et al.*, 2011). Recent implementations have incorporated more advanced models such as the birth-death model (Stadler, 2009), Susceptible-Infected-Recovered (SIR) population model (Rasmussen *et al.*, 2011) and including infection heterogeneity for an HIV dataset (Frost & Volz, 2013). While progress has been made to allow more realistic scenarios of spatial spread in phylodynamic space (Lemey *et al.*, 2010) incorporating landscape heterogeneity is still at a developmental, experimental stage (Chapter 4). In addition, there are problems with the effect of both spatial and temporal sampling bias on inference that still haven't been resolved (Frost *et al.*, 2015). We applied extensions to existing Bayesian frameworks to explicitly incorporate landscape heterogeneity by modifying the spatial landscape in both a discretised and continuous state. Delimiting population structure to allow discrete-state analyses does not naturally apply to systems characterised by continuous diffusion and presents problems relating to spatial scale and interpretation that make a generalised discrete approach difficult. Continuous diffusion models are a more biologically realistic presentation of disease diffusion in most scenarios and further development of our approach using multidimensional scaling could generate a useful phylodynamic framework. However, as discussed in Chapter 4 a better way to formalise and interpret results is necessary, for which simulation models could prove useful.

Aside from directly incorporating landscape heterogeneity, as we explored in Chapter 4, post-processing analyses can potentially provide insightful information in the interim between testing and finalising spatially heterogeneous phylogeographic models. Diffusion rate variation across branches in reconstructed spatiotemporally-referenced phylogenies provides a useful metric to explore correlations between landscape variables. This approach has recently been used by Dellicour *et al.* (2015) in an R package called Seraphim, which provides functions to extract branch-specific rates, estimate summary statistics and examine correlations between dispersal rates and environmental variables. Importantly, the means to test significance against null models is included, including comparison to a torus translation and reflection of the original landscape, which keeps spatial correlation intact. Applying these methods to the data presented here would be an obvious next step to complement results so far.

The second approach, demonstrated in Chapter 5, uses models of transmission to explicitly account for host population structure (e.g. SEIR models) and the underlying epidemiological processes, achieving the best approximation of "who infected whom" strengthened by inferences from genetic data. Recent statistical frameworks have used powerful Bayesian inference schemes that synergistically infer transmission trees from genetic and epidemiological

data (Hall *et al.*, 2015; Jombart *et al.*, 2014; Mollentze *et al.*, 2014a; Morelli *et al.*, 2012; Ypma *et al.*, 2012, 2013). These have been used to estimate key epidemiological parameters such as the rate of mutation per transmission event (Cottam *et al.*, 2008), the proportion of unobserved cases (Mollentze *et al.*, 2013), the effective reproductive rate (Jombart *et al.*, 2014) and most likely transmission pathways (Jombart *et al.*, 2014; Morelli *et al.*, 2012; Ypma *et al.*, 2012). Although these inference methods are relatively new, progress has been swift and shows much promise for the future. Specifically, the processes mentioned above could be directly incorporated in algorithms to improve inference and post-hoc analyses could provide additional means to detect landscape level effects. In Chapter 5 some observational measures of correlation between landscape heterogeneity and transmission events were made but more robust, quantifiable tests could be utilised to explore this more thoroughly. As the nature of questions surrounding transmission biology evolves so does the potential to develop these methods, including the generation of a truly synthetic framework. Currently this is an area very much at the forefront of understanding fundamental processes in infectious disease biology but increasingly available high resolution data is contributing to progress.

6.0.5 Beyond the consensus

One of the obvious limitations of the work presented in this thesis is the use of consensus sequences to represent individuals. This level of characterisation may be adequate for population level epidemiological inference but many key evolutionary processes occur beyond the consensus (Holmes & Grenfell, 2009). RNA viruses exist as complex, heterogeneous within-host populations where consensus sequences represent the dominant sequence in the population (Holmes & Moya, 2002), essentially ignoring the diverse sub-structure of minority variants. This additional measure of diversity may provide information to further delineate single host-to-host transmission events and provide a better understanding of the processes that determine how genetic diversity is transmitted between hosts and different evolutionary scales (Morelli *et al.*, 2013; Pybus & Rambaut, 2009).

Research on foot and mouth virus populations indicates that the rate of nucleotide substitution between hosts is faster than within-host rates (Orton *et al.*, 2013), while the opposite appears to be true in HIV (Lemey *et al.*, 2006). In the case of foot and mouth this suggests that population bottlenecks between hosts influence the fixation rate of mutations in consensus level sequences (Orton *et al.*, 2013). As rabies is a multi-host pathogen it would be interesting to characterise what effect transmission bottlenecks have in different transmission chains including dog-to-dog and cross-species transmissions, for example between livestock and dogs. Other questions include what selective pressures exist within the host during the incubation period and what effect the duration of the incubation has on within-host diversity, and if convergent evolution occurs in different hosts (Mollentze *et al.*, 2014a; Pybus & Rambaut, 2009).

6.0.6 Implications for control and surveillance

Despite it's notoriety as a fatal disease with terrible clinical manifestations, rabies falls into the category of neglected tropical diseases. Canine rabies can be successfully controlled through mass dog vaccination and has been eliminated or is near elimination in many areas of the world (King *et al.*, 2004; Velasco-Villa *et al.*, 2008; Vigilato *et al.*, 2013). However, it remains a persistent threat in many developing countries where limited resources and cultural challenges allow the disease to be maintained.

WHO recommended vaccination coverage to control rabies is $\geq$70% (WHO, 2013). However, considerable variation in the observed levels of coverage that have been both successful and unsuccessful in controlling canine rabies suggests there is some inherent heterogeneity that could be utilised to optimise vaccination strategies. Determining the source of this variation is challenging, illustrated by our analyses in Chapter 4. We found evidence to implicate several landscape features as mediators of dispersal but patterns were typically noisy and hard to fully characterise. However, the potential role of landscape features in structuring rabies events suggests that there way be ways to strengthen and inform vaccination campaigns by exploiting this pre-existing knowledge. Rivers achieved the highest levels of support in our phylodynamic based approach and also appeared to limit transmission events in transmission tree reconstructions in Chapter 5. Rivers have previously been recognised as barriers in wildlife rabies systems, reducing the dispersal of fox rabies in Europe (Wandeler *et al.*, 1988) and raccoon rabies at township level in the United States (Smith *et al.*, 2002), but this is the first time the effect has been shown for dog rabies beyond a global scale. The effect of features as barriers can be further tested through simulation studies.

Studies using genetic information in transmission tree reconstructions have been based on the assumption that host to host transmission is genetically homogeneous. However, individual heterogeneity and the influence of external factors may account for variation in transmission that could have important consequences for overall disease dynamics (Lloyd-Smith *et al.*, 2005). Hampson *et al.* (2009) highlighted individual heterogeneity in RABV transmission in the Serengeti, with some individuals causing significantly more secondary infections than the population estimates of R_0. The effects of individual heterogeneity may overpower underlying landscape structure, essentially masking patterns associated with landscape processes. Dogs with the propensity to bite more than normal i.e. superspreaders may still be influenced by certain aspects of landscape heterogeneity, particularly physical features such as rivers or roads. This might be the reason why we found the greatest support for this type of landscape feature as predictors of diffusion (rivers, roads and slope were the most supported landscape predictors in Chapter 4). It may be insightful to characterise spatial patterns according to normal versus superspreading individuals, but this relies on the assumption that all superspreading events are captured by contact tracing. Alternatively, if superspreaders are not known *a priori* genetic data can be used to identify superspreader dynamics using

phylogenetic methods (Stadler & Bonhoeffer, 2013) or transmission tree reconstructions as in Chapter 5.

This thesis highlights many of the practical considerations of using genetic data to characterise infectious disease dynamics, particularly in areas where significant logistical challenges exist. We experienced challenges from the starting point of obtaining genetic material from field samples of varying quality to how genetic sampling ultimately affects the inference of infectious disease dynamics. Samples collected in the field often vary in quality and quantity, particularly when facilities for sample storage are problematic (e.g. power cuts to freezers) or samples need to be collected and transported from rural areas. Hence the state of genetic material present in starting material is not always ideal and presents a significant challenge in terms of ultimately finding and extracting material for next generation sequencing protocols. That we were able to extract and sequence genetic material from a large number of samples (179 whole genomes in total) is encouraging evidence that this level of genetic characterisation can be achieved in less than optimal circumstances.

One of the most interesting findings in this research was the identification of several co-circulating lineages in the Serengeti District, suggesting a high level of diversity at a very small scale. However, the inference from genetic characterisation was limited by the spatial extent of genetic sampling. The Serengeti was over-represented in the sample set for phylodynamic reconstruction and neighbouring areas were not sampled. It is possible that the genetic diversity observed in the Serengeti District has a wider distribution reflecting a larger scale of transmission. Essentially, we are missing information on an important spatial scale in the hierarchy of transmission dynamics, including information regarding the influx of virus from other areas and the wider distribution of viral lineages. This information has important implications on recommendations for control . Based on our results it seems pertinent to recommend further whole genome characterisation of rabies at this larger spatial scale before a consensus can be achieved on the best strategy for effective control. Once this has been achieved genetic surveillance at this level may be most informative and best utilised in the later stages of control when it is imperative to quickly regain control when an outbreak occurs and identify sources of incursion.

Overall, this thesis has highlighted the potential for future landscape studies to characterise transmission and provides a framework for further development. When inference methods have been formalised and robust summary statistics can be generated landscape epidemiology studies have the potential to generate data that can directly inform in control efforts, such as the permeability of existing barriers to rabies dispersal.

APPENDIX A

Chapter 2 Appendix

Table A.1: GenBank accession numbers and details of rabies virus whole genome sequences used in a global phylogenetic reconstruction for Chapter 2

Accession no	Country	Year	Host
AB362483	Brazil	2002	Fox
AB517659	Brazil	na	Domestic dog
AB517660	Brazil	na	Fox
AB569299	Sri Lanka	2008	Human
AB635373	Sri Lanka	2009	Golden palm civet
AY956319	India	2004	Human
EF437215	India	na	Human
EU293111	Thailand	1983	Human
EU293115	France	1991	Fox
EU293121	Thailand	1983	Human
EU643590	China	2006	Human
FJ712193	China	2008	Domestic dog
FJ712194	China	2008	Domestic dog
FJ712195	China	2008	Ferret
FJ712196	China	2008	Ferret
FJ866835	China	2008	Domestic dog
FJ866836	China	2008	Domestic dog
GU345746	China	1992	Domestic dog
GU345747	China	1986	Human
GU345748	China	2006	Domestic dog
GU647092	China	2008	Chinese ferret badger
HQ317918	China	1956	Human
HQ450385	China	2008	Domestic dog
HQ450386	Mexico	na	Domestic dog
JN609295	China	2008	Domestic dog
JN786877	Thailand	na	Domestic dog
JQ423952	China	2011	Horse
JQ647510	China	2011	Donkey
JQ685894	USA	1994	Striped skunk
JQ685899	USA	2009	Gray fox
JQ685943	USA	2009	Gray fox
JQ685944	USA	1984	Striped skunk
JQ685967	USA	na	Striped skunk
JQ685970	USA	1974	Striped skunk
JQ685975	Mexico	2009	Spotted skunk
JQ730682	China	2010	Domestic dog
JQ944704	Russia	2009	Raccoon dog
JQ944705	Russia	2008	Domestic dog
JQ944706	Russia	2008	Domestic dog
JQ944707	Russia	2008	Deer
JQ944708	Russia	2008	Red fox
JX088694	China	na	Pig

JX473838	Namibia	2009	Jackal
JX473839	Namibia	2009	Jackal
JX473840	Namibia	2009	Namibian kudu
JX473841	Namibia	2009	Namibian kudu
KC169986	China	2009	Cow
KC171643	South Korea	2008	Cow
KC171644	South Korea	2004	Raccoon dog
KC171645	South Korea	1999	Raccoon dog
KC193267	China	2011	Cow
KC196743	Nigeria	2011	Domestic dog
KC737850	USA	2011	Human
KC762941	China	2009	Ferret badger
KF154996	India	1987	Human
KF154997	Estonia	na	Raccoon dog
KF154998	Israel	1950	Domestic dog
KF154999	SrinaLanka	2008	Domestic dog
KF155000	Iraq	2010	Cow
KF155001	Morocco	2009	Cow
KF155002	Tanzania	2010	Domestic dog

APPENDIX B

Chapter 3 Appendix

B.1 Supplementary material

B.1.1 Partial nucleoprotein (405bp) sequences

An additional 50 partial N gene sequences were generated from samples obtained from Tanzania between 2004 and 2013 (Table S2). cDNA was prepared as described in the main text and a 405bp fragment amplified using hemi-nested PCR incorporating pan-Lyssavirus primers JW6UNI for first round products or JW10UNI for second round products, in combination with JW12, as previously described (Heaton *et al.*, 1997). Products were visualised on a 2% agarose gel with SYBR Safe DNA Gel Stain (Invitrogen). First round positive PCR products were purified using a QIAquick PCR purification kit (Qiagen) and approximately 50-150ng of product was used in a sequencing reaction with the Big Dye sequencing kit (Applied Biosystems). Sequencing was performed with the same primers on an ABI 3100 machine at the APHA sequencing facility and a consensus generated as previously described (Heaton *et al.*, 1997).

B.1.2 Partial genome datasets

Box B.1: GenBank search for >400bp Rabies virus sequences from Africa
((tanzania) OR africa) OR south africa) OR mozambique) OR zimbabwe) OR MAD) OR malawi) OR botswana) OR namibia) OR zambia) OR angola) OR congo) OR burundi) OR democratic republic of congo) OR burundi) OR rwanda) OR uganda) OR cameroon) OR gabon) OR central african republic) OR sudan) OR MAD[Text Word]) OR MOZ[Text Word]) OR NGA[Text Word]) OR zaire) OR madagascar) OR CAR) OR Mozambique) OR CAF) OR kenya) OR somalia) OR ethiopia) OR chad) OR nigeria) OR Nigeria) OR benin) OR togo) OR ghana) OR niger) OR mali) OR liberia) OR sierra leone) OR guinea) OR burkina faso) OR libya) OR mauritania) OR algeria) OR eritrea) OR egypt) OR morocco) OR tunisia) OR senegal) OR tanzania) OR Djibouti) OR Cote d'Ivoire) OR somalia) OR Guinea-Bissau) OR gambia) OR Republic of the Congo) OR Congo) OR Equatorial Guinea) OR ivory coast) OR swaziland) OR lesotho) OR Kissi[Author]) OR Talbi[Author]) AND rabies virus[Organism]) AND nucleoprotein[Title]) AND 400:11930[Sequence Length]) NOT turkey) NOT lebanon) NOT israel) NOT jordan)))))))))) NOT rabies virus strain[Title])))))

Nucleoprotein sequences were retrieved from GenBank by searching for rabies virus records containing text matching Africa or any country in Africa. Only sequences with a length greater than 400bp were accepted and vaccines strains were excluded from the search. In addition, several datasets known to exist but which did not contain searchable text references were manually added to the search criteria (Box B.1). Fasta files and GenBank records were downloaded in March 2015 and filtered to remove isolates not from Africa.

B.1.3 Additional whole genome sequencing protocols

During the optimisation of protocols for whole genome sequencing some sequences used in the final analysis were generated by the following methods:

i) Amplicon sequencing: 454 platform
To generate cDNA μl of TRIzol-extracted viral RNA was subject to reverse transcription using the

primer RABV_Tzdg.p1f (5' ACGCTTAACAACAAAATCAGAG 3') at a concentration of 2pmolμl and Superscript III reverse transcriptase (Invitrogen) in a total volume of 20mul, as per manufacturer's instructions. A working set of 26 short, tagged, overlapping primer pairs spanning the entire RABV genome was designed based on the full length Tanzanian dog reference genome RV2772 (accession: KF155002). Primers were used to obtain PCR products with 1μl of cDNA and a KOD hot start DNA polymerase kit as per manufacturer's protocol (Novagen) with the following cycling parameters: 1 hold at 95
newunicodecharřDegree of *Doctor of Philosophy* (Ph.D.)C for 2 mins, 35 cycles at 95C for 20 s, 50-60C (dependent on the optimised temperature for each primer pair) for 20 s, 70C for 20 s and a final hold at 70Cfor 10 min using a 2720 thermal cycler (Applied Biosystems). Products were pooled for each sample and sent to the APHA sequencing facility for template preparation and 454 pyrosequencing.

ii) Depleted RNA: 454 platform
TRIzol-extracted viral RNA was depleted of host genomic DNA and ribosomal RNA as described in the main text. Random-primed cDNA libraries were constructed and sent to the APHA sequencing facility for 454 pyrosequencing.

iii) Depleted RNA: Illumina NextSeq 500 platform
Double stranded cDNA was synthesized as described in the main text and sent to the Glasgow Polyomics facility (University of Glasgow, Glasgow, UK) for Nextera XT library preparation and sequencing on an Illumina NextSeq 500 platform.

Table B.1: Statistics for the total number of rabies virus samples used in this thesis showing the number of PCR positive samples obtained from suspect cases sent from Tanzania and the success rate for obtaining consensus level NGS data from prepared sequence libraries.

Shipment	No of samples	PCR positive	% Positive	No of libraries prepared	Successful NGS	% Sequenced successfully
2009	61	22	0.36	18	16	0.89
May 2011	52	31	0.60	48	43	0.90
Dec 2011	92	70	0.76	47	41	0.87
2012	99	67	0.68	65	52	0.80
2013	69	42	0.61	42	33	0.79
			Totals	220	185	0.84

Table B.2: Epidemiological information and whole genome sequencing (WGS) details for Tanzanian whole genome samples used in Chapter 3 (*reference sequence).

Region	Date	Easting	Northing	Species	WGS protocol	Sample ID	% Genome coverage	Average depth of coverage	Accession no
Arusha	18-Apr-04	777032	9793554	Cow	MiSeq	RV2502	100	165	KR906739
Arusha	18-Jun-03	784690	9771564	Domestic dog	MiSeq	RV2504	100	1871	KR906741
Dar es Salaam	03-May-10	1207114	9234522	Domestic dog	MiSeq	RV2770	100	22	KR906743
Dar es Salaam	02-Nov-10	1182523	9234308	Domestic dog	MiSeq	RV2774	100	146	KR906746
Iringa	27-May-10	750553	9053269	Domestic dog	MiSeq	RV2771	100	42	KR906744
Iringa	26-Nov-10	750553	9053269	Domestic dog	MiSeq	RV2775	100	120	KR906747
Lindi	07-Feb-11	1088355	8845396	NA	Depleted RNA: 454	RV2780	100	20	KR906751
Lindi	28-Feb-11	1088355	8845396	NA	MiSeq	RV2784	100	27	KR906754
Lindi	03-Oct-10	1088355	8845396	NA	Depleted RNA: 454	RV2807	100	10	KR906757
Morogoro	06-Aug-08	244890	9075488	Domestic dog	MiSeq	RV2498	100	61	KR906735
Morogoro	05-Aug-08	243727	9075748	Domestic dog	MiSeq	RV2499	100	528	KR906736
Morogoro	04-Apr-10	1040336	9253119	Domestic dog	MiSeq/ NextSeq	RV2808	99	28	KR906758
Morogoro	16-Apr-10	1032182	9228490	Domestic dog	MiSeq	RV2809	98	20	KR906759

Morogoro	23-Apr-10	1065994	9258751	Domestic dog	MiSeq	RV2810	100	101	KR906760
Morogoro	24-Apr-10	1033809	9250464	Domestic dog	MiSeq/ NextSeq	RV2811	98	16	KR906761
Morogoro	28-Apr-10	352510	9245700	NA	MiSeq	RV2813	100	36	KR906762
Morogoro	17-May-10	898066	8996235	Pig	MiSeq	RV2814	100	962	KR906763
Morogoro	27-Sep-10	9105076	242926	Cow	MiSeq	RV2815	100	77	KR906764
Morogoro	28-Sep-10	242926	9105076	Cow	MiSeq	RV2816	100	117	KR906765
Mtwara	28-Feb-11	1193634	8807598	NA	MiSeq	RV2783	100	153	KR906753
Pemba	29-Dec-12	1256936	9414352	Domestic dog	MiSeq	RV2776	100	499	KR906748
Pemba	24-Dec-12	1256936	9414352	Domestic dog	MiSeq/ NextSeq	RV2777	100	57	KR906749
Pemba	26-Dec-12	1256936	9414352	Domestic dog	MiSeq	RV2778	99	41	KR906750
Pemba	11-Feb-11	1205189	9313425	Domestic dog	MiSeq	RV2782	100	226	KR906752
Pemba	26-Nov-10	1256936	9414352	Domestic dog	MiSeq	RV2817	100	44	KR906766
Pwani	02-Aug-10	1138047	9246282	Domestic dog	Genbank	RV2772	na	na	KF155002*
Pwani	26-Oct-10	1138047	9246282	Domestic dog	MiSeq/ NextSeq	RV2773	100	104	KR906745
Serengeti	12-Jul-08	696385	9802796	Domestic dog	MiSeq	RV2495	100	129	KR906734
Serengeti	23-Nov-08	698019	9804712	Domestic dog	MiSeq	RV2500	100	16	KR906737
Serengeti	14-Feb-04	679372	9810508	Domestic dog	MiSeq	RV2501	100	108	KR906738

Serengeti	15-Feb-09	686729	9761151	Wild Cat	MiSeq	RV2503	100	29	KR906740
Serengeti	12-Sep-09	665178	9777064	Domestic dog	MiSeq	RV2767	100	250	KR906742
Serengeti	02-Jan-11	657176	9822908	Cow	MiSeq	RV2793	100	39	KR906755
Serengeti	29-Jan-11	680129	9805797	Domestic dog	MiSeq	RV2799	100	41	KR906756
Serengeti	11-May-11	691423	9791388	Domestic dog	MiSeq	RV2861	100	1259	KR906767
Serengeti	11-May-11	674259	9805687	Domestic dog	MiSeq	RV2862	100	21	KR906768
Serengeti	17-Jun-11	701315	9794641	Domestic dog	Amplicon seq: 454	RV2871	100	25	KR906769
Serengeti	29-Jun-11	680196	9811835	Domestic dog	MiSeq	RV2875	100	99	KR906770
Serengeti	15-Aug-11	653024	9822353	Domestic dog	MiSeq	RV2894	100	1470	KR906771
Serengeti	19-Aug-11	681869	9798035	Domestic dog	MiSeq	RV2896	100	69	KR906772
Serengeti	27-Sep-11	656396	9803751	Domestic dog	MiSeq	RV2900	100	84	KR906773
Serengeti	22-Sep-11	653179	9802910	Domestic dog	MiSeq	RV2901	100	107	KR906774
Serengeti	24-Sep-11	684532	9790009	Domestic dog	MiSeq	RV2902	99	44	KR906775
Serengeti	16-Oct-11	700560	9803728	Cow	MiSeq	RV2907	100	71	KR906776
Serengeti	05-Dec-11	696698	9792023	Domestic dog	MiSeq/ NextSeq	RV3091	100	497	KR906777
Serengeti	22-Dec-11	681720	9798404	Domestic dog	MiSeq	RV3093	100	43	KR906778

Serengeti	19-Feb-12	698140	9804256	Domestic dog	MiSeq	RV3100	100	24	KR906779
Serengeti	22-Dec-11	648913	9823033	Sheep	MiSeq	RV3104	100	95	KR906780
Serengeti	09-Apr-12	656704	9803322	Domestic dog	MiSeq	RV3107	100	293	KR906781
Serengeti	29-Apr-12	650984	9807685	Domestic dog	MiSeq	RV3111	100	27	KR906782
Serengeti	12-May-12	669765	9795977	Domestic dog	MiSeq	RV3117	100	58	KR906783
Serengeti	26-Apr-12	647442	9799177	Domestic dog	MiSeq	RV3123	100	65	KR906792
Serengeti	06-Jun-12	658003	9809433	Goat	MiSeq/ NextSeq	RV3125	100	106	KR906784
Serengeti	11-Jun-12	681576	9797817	Domestic dog	MiSeq	RV3127	100	204	KR906785
Serengeti	16-Jun-12	674458	9797076	donkey	MiSeq	RV3128	100	89	KR906786
Serengeti	07-Jul-12	685498	9797037	Domestic dog	MiSeq	RV3131	100	113	KR906787
Serengeti	02-Apr-12	656028	9801566	Domestic dog	MiSeq	RV3132	100	14	KR906788
Serengeti	04-Jul-12	700856	9800221	Domestic dog	MiSeq/ NextSeq	RV3133	100	118	KR906789
Serengeti	27-Jul-12	656945	9804374	Domestic dog	MiSeq	RV3140	100	738	KR906790
Serengeti	23-Jan-12	657742	9809576	Domestic dog	MiSeq	RV3149	93	6	KR906791

Table B.3: Epidemiological information and whole genome sequencing (WGS) details for Tanzanian whole genome samples used in Chapter 3 (*reference sequence).

Sample	Region	Species	Year	Accession.no
RV2490	Serengeti	Domestic dog	2008	KR534217
RV2491	Serengeti	Domestic dog	2008	KR534218
RV2492	Serengeti	Domestic dog	2007	KR534219
RV2493	Serengeti	Domestic dog	2008	KR534220
RV2494	Serengeti	Cat	2009	KR534221
RV2496	Serengeti	Honey badger	2004	KR534222
RV2497	Serengeti	Domestic dog	2007	KR534223
RV2768	na	Domestic dog	na	KR534224
RV2769	Iringa	Domestic dog	2010	KR534225
RV2779	Pwani	Domestic dog	2011	KR534226
RV2787	Serengeti	Domestic dog	2010	KR534227
RV2788	Serengeti	Domestic dog	2010	KR534228
RV2789	Serengeti	Domestic dog	2010	KR534229
RV2790	Serengeti	Domestic dog	2010	KR534230
RV2791	Serengeti	Domestic dog	2010	KR534231
RV2792	Serengeti	Jackal	2011	KR534232
RV2794	Serengeti	Domestic dog	2011	KR534233
RV2795	Serengeti	Cow	2011	KR534234
RV2796	Serengeti	Cow	2011	KR534235
RV2797	Serengeti	Cow	2011	KR534236
RV2798	Serengeti	Cow	2011	KR534237
RV2800	Serengeti	Domestic dog	2011	KR534238
RV2802	na	Na	2011	KR534239
RV2804	na	Na	na	KR534240
RV2806	na	Na	na	KR534241
RV2856	na	Na	na	KR534242
RV2857	Serengeti	Domestic dog	2011	KR534243
RV2889	Serengeti	Domestic dog	2011	KR534244
RV2890	Serengeti	Cow	2011	KR534245
RV2891	Serengeti	Civet	2011	KR534246
RV2892	Serengeti	Domestic dog	2011	KR534247
RV2893	Serengeti	Cow	2011	KR534248
RV2895	Serengeti	Domestic dog	2011	KR534249
RV2897	Serengeti	Domestic dog	2011	KR534250
RV2898	Serengeti	Domestic dog	2011	KR534251
RV2899	Serengeti	Domestic dog	2011	KR534252
RV2903	Serengeti	Domestic dog	2011	KR534253
RV2906	Serengeti	Cat	2011	KR534254
RV2908	Serengeti	Goat	2011	KR534255
RV2909	Serengeti	Cow	2011	KR534256
RV2910	Morogoro	Na	2011	KR534257
RV2911	Morogoro	Na	2011	KR534258

RV2913	Morogoro	Na	2011	KR534259
RV2914	Morogoro	Na	2011	KR534260
RV2915	Morogoro	Na	2011	KR534261
RV2916	Morogoro	Na	2011	KR534262
RV2917	Morogoro	Na	2011	KR534263
RV2920	Morogoro	Na	2011	KR534264
RV2921	Morogoro	Na	2011	KR534265
RV2922	na	Na	2011	KR534266

Table B.4: Model comparisons for molecular clock models from marginal likelihood estimates using path sampling (PS) and stepping stone (SS) sampling in BEAST v1.8.1

Molecular clock model	Migration rate prior	PS	SS
Strict	HKY	-32450.10	32451.94
	GTR	-32258.12	-32262.52
Relaxed uncorrelated lognormal	HKY	-32333.16	-32336.84
	GTR	**-32205.06**	**-32208.48**
Relaxed uncorrelated exponential	HKY	-32341.10	-32345.95
	GTR	-32206.50	-32210.39

Table B.5: Model comparison between gene-specific or gene-linked HKY or GTR nucleotide models with different codon position partitioned models (alignment has 5 genes and 1 concatenated non-coding sequence partition). Marginal likelihood estimation using path sampling (PS) and stepping stone (SS) sampling in BEAST v1.8.1 was used for model selection. The best model is indicated in bold.

Model description	Codon partition model	Number of partitions	Model	PS	SS
Gene partitioned with gene-specific rate variation	NA	6	5 genes * (HKY + Γ) + Non-coding (HKY + Γ)	-31792.27	-31794.86
Gene linked & codon position partitioned: among codon position rate heterogeneity, homogeneous rates among genes	CP112	2	Gene linked (HKY112 + CP112 + Γ112) + Non-coding (HKY + Γ)	-30955.51	-30958.76
	CP123	2	Gene linked (HKY112 + CP112 + Γ112) + Non-coding (HKY + Γ)	-30888.21	-30891.53
Gene partitioned & codon position partitioned: among codon position rate heterogeneity, heterogeneous rates among genes	CP112	6	5 genes * (HKY112 + CP112 + Γ112) + Non-coding (HKY + Γ)	-31032.33	-31037.16
	CP123	6	5 genes * (HKY112 + CP123 + Γ123) + Non-coding (HKY + Γ)	-30969.91	-30975.99
Gene partitioned with gene-specific rate variation	NA	6	5 genes * (GTR + Γ) + Non-coding (GTR + Γ)	-31776.52	-31780.77

Gene linked & codon position partitioned: among codon position rate heterogeneity, homogeneous rates among genes	CP112	2	Gene linked (GTR112 + CP112 + Γ112) + Non-coding (GTR + Γ)	-30920.62	-30924.50
	CP123	**2**	**Gene linked (GTR123 + CP123 + Γ123) + Non-coding (GTR + Γ)**	**-30831.04**	**-30835.69**
Gene partitioned & codon position partitioned: among codon position rate heterogeneity, heterogeneous rates among genes	CP112	6	5 genes * (GTR112 + CP112 + Γ112) + Non-coding (GTR + Γ)	-31156.41	-31163.37
	CP123	6	5 genes * (GTR112 + CP123 + Γ123) + Non-coding (GTR + Γ)	-31075.10	-31079.98

Table B.6: Model comparisons for different migration rate priors from marginal likelihood estimates using path sampling (PS) and stepping stone (SS) sampling in BEAST v1.8.1

Migration rate prior	PS	SS
Exponential mean=0.001	-30567.84	-30568.29
Exponential mean=0.01	**-25028.91**	**-25029.04**
Exponential mean=0.1	-30688.06	-30688.37
Exponential mean=0.5	-30867.20	-30868.62
Exponential mean=1	-30817.55	-30815.80

Table B.7: Bayes Factor (BF) support for significant rabies virus diffusion pathways in Tanzania identified under a BSSVS procedure and median (with range) number of transitions along those pathways (shown with posterior probability of transition occurring in the phylogeny) estimated via Markov jump counts in BEAST.

From	To	BF	Transitions	Posterior probability of transition
Serengeti	Morogoro	135.302	6 (3-10)	1
Arusha	Mtwara	7.39	1 (1-3)	0.72
Dar	Iringa	5.90	1 (1-2)	0.39
Pwani	Serengeti	5.22	1 (1-5)	0.05
Iringa	Arusha	4.59	1 (1-2)	0.27
Mtwara	Arusha	4.56	1 (1-3)	0.2
Arusha	Iringa	4.31	1 (1-3)	0.61
Pemba	Lindi	4.24	1 (1-3)	0.19
Morogoro	Dar	4.23	1 (1-2)	0.05
Pwani	Pemba	4.05	1 (1-2)	0.07
Iringa	Dar	4.00	1 (1-3)	0.37
Pwani	Lindi	3.45	1 (1-2)	0.04
Pemba	Serengeti	3.37	1 (1-6)	0.11
Lindi	Pemba	3.24	1 (1-4)	0.18
Dar	Lindi	3.21	1 (1-4)	0.08
Morogoro	Pemba	3.16	1 (1-4)	0.53
Pemba	Pwani	3.07	1 (1-2)	0.04
Morogoro	Pwani	3.07	1 (1-2)	0.06

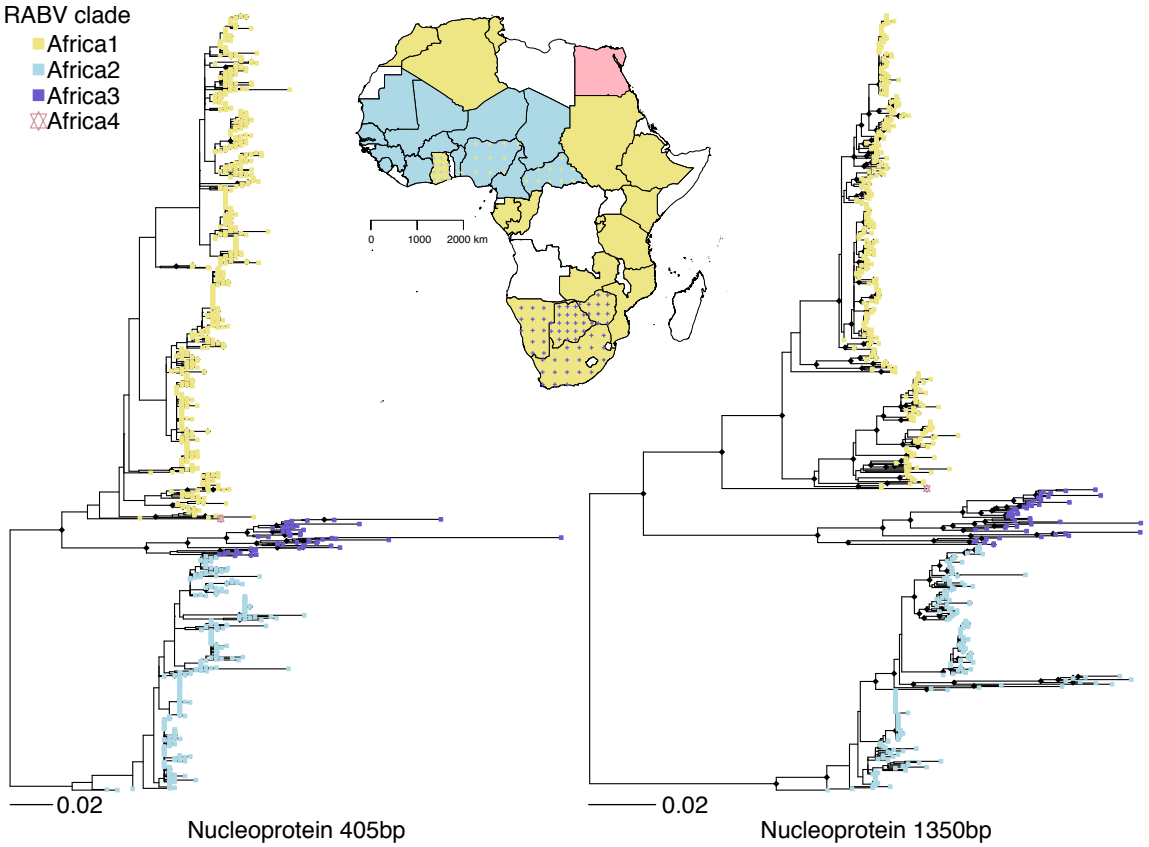


Figure B.1: Maximum likelihood trees derived from datasets of rabies virus (RABV) sequences from Africa for a) a 405bp fragment of the nucleoprotein (N) gene (n=1317) and b) full length 1350bp nucleoprotein gene sequences (n=674). Samples are colour-coded according to major RABV clades in Africa and their spatial distribution indicated on the map. Countries in which more than one clade was sampled have coloured crosses to indicate the less frequently sampled clade. Trees are scaled by number of substitutions per site.

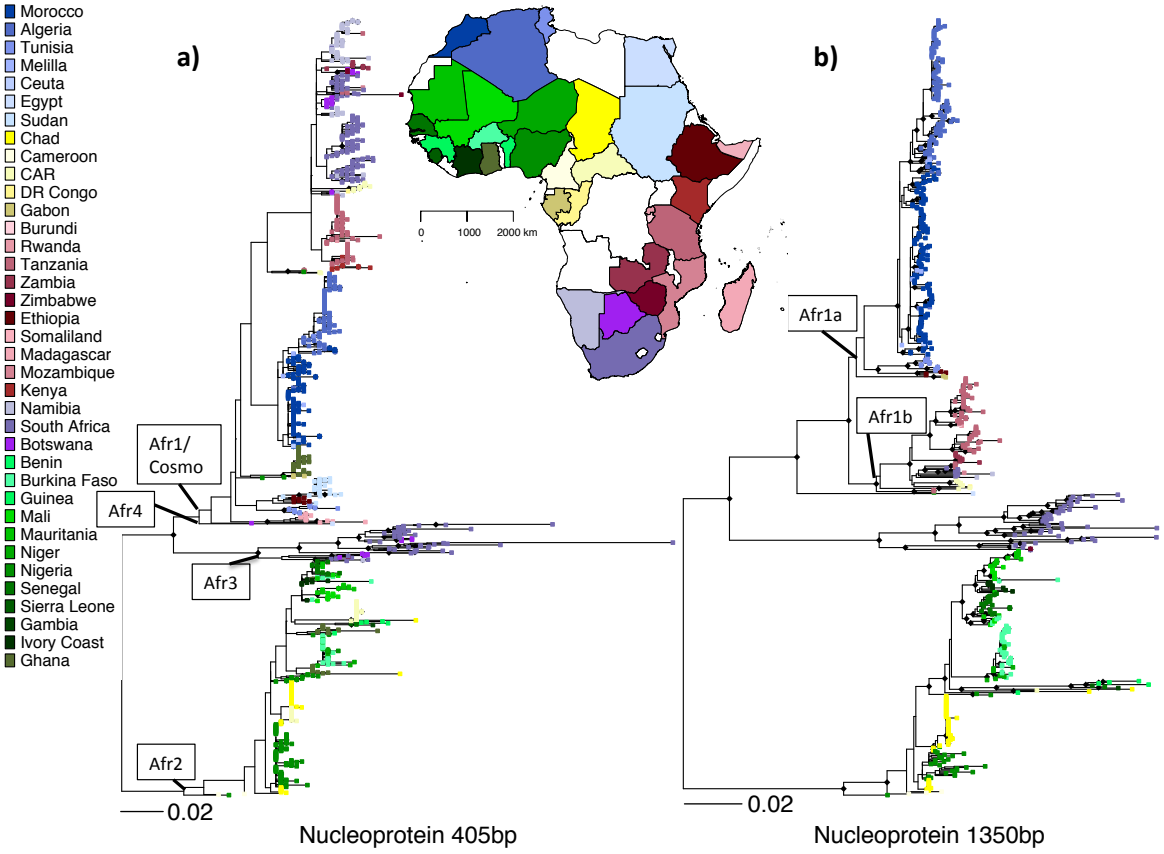


Figure B.2: Maximum likelihood trees derived from datasets of rabies virus sequences from Africa for a) a 405bp fragment of the nucleoprotein (N) gene (n=1397) with major African RABV clades indicated (Afr1/Cosmo: Africa 1/Cosmopolitan, Afr2: Africa2; Afr3: Africa 3, mongoose-associated clade; Afr4: Africa 4); and b) full length 1350bp nucleoprotein gene sequences (n=769) with the two Africa 1 subclades shown. Samples are coloured according to their country of origin as indicated on the map. All countries were sampled to at least partial N resolution. Trees are scaled by number of substitutions per site.

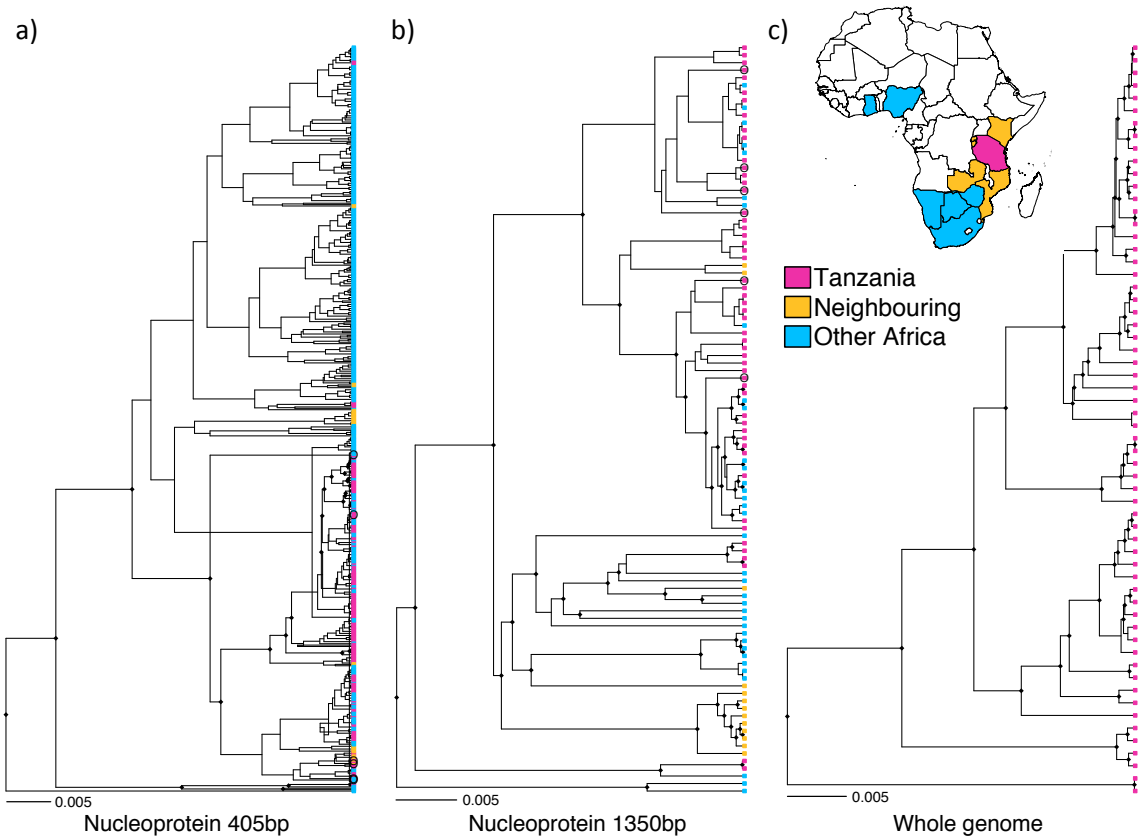
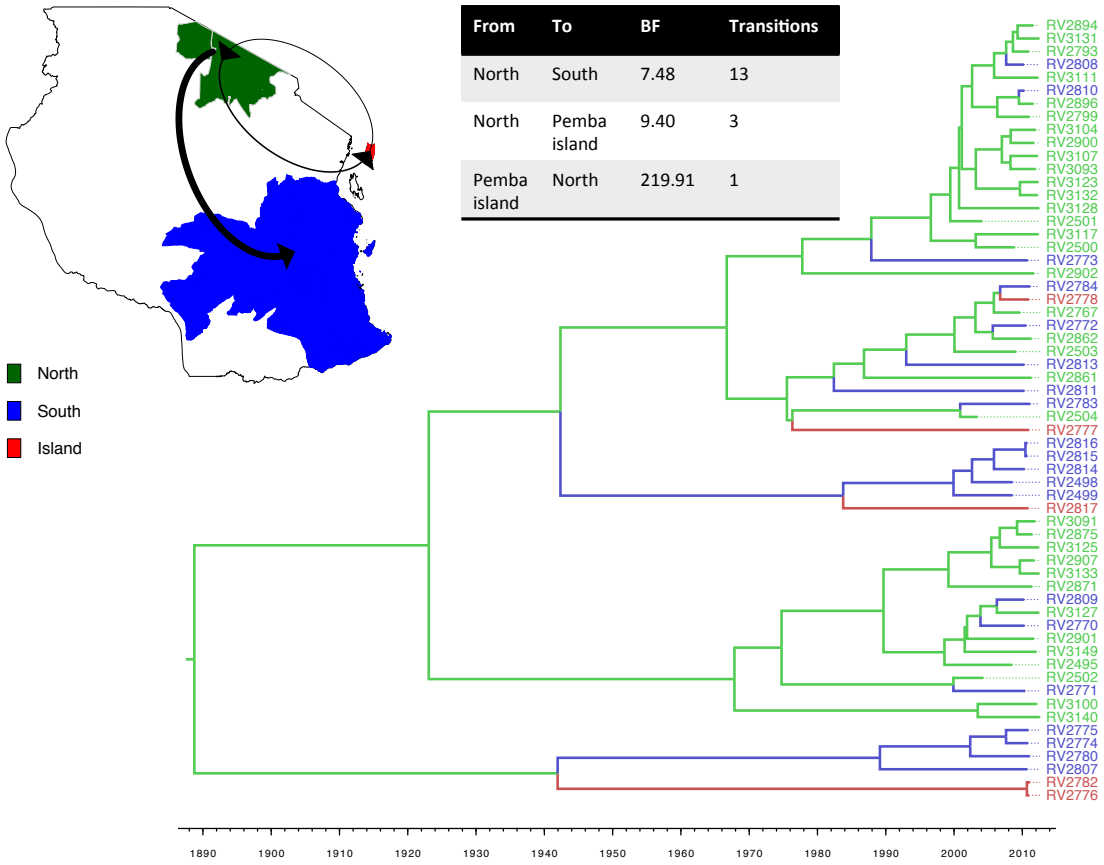


Figure B.3: Maximum clade credibility trees from Bayesian phylogenetic estimation in BEAST for datasets of rabies virus sequences from the Africa 1B clade for increasing levels of genome coverage: a) a 405bp fragment of the nucleoprotein gene (n=510) from countries highlighted on the map, b) full 1350bp nucleoprotein gene (n=100) from the same countries except Botswana, Ghana, Kenya and Zimbabwe; and c) whole genome sequences from Tanzania. Trees are scaled by number of substitutions per site and diamonds indicate nodes with posterior probability support $\geq$0.9. Older samples from the Serengeti District (~20years old) are circled in the partial genome trees.



From	To	BF	Transitions
North	South	7.48	13
North	Pemba island	9.40	3
Pemba island	North	219.91	1

Figure B.4: North-south phylogeographic structure among 60 rabies virus whole genome sequences isolated in Tanzania from 2003 to 2012. A maximum clade credibility tree is shown with branches coloured according to the most probable posterior location of their descendent nodes, inferred by discrete-state phylogeographic reconstruction using BEAST. The tree is scaled according to time in years and diamonds indicate node posterior support $\geq$0.9. The map and key indicate spatial division according to locations in the northern mainland (n=35), southern mainland (n=20) or Pemba island (n=5). Inset table provides details of dispersal pathways with Bayes Factor results and the estimated number of transitions according to Markov jumps counts, shown on the map with arrow width scaled by the number of transitions.

APPENDIX C

Chapter 4 Appendix

Table C.1: Epidemiological information and whole genome sequencing (WGS) details for 152 whole genome samples sampled from the Serengeti District in Tanzania between 2004 and 2013. Samples used contained in the window used for space-time-genetic inference in Chapter 5 have an asterisk.

Sample	Date of infection	NGS protocol	Total reads	Reads mapped	Proportion mapped	Average depth	Easting	Northing	Host
RV2483	28/09/08	Miseq	397221	2045	0.51	22.05	701590	9794710	Goat
RV2485	28/09/08	2*Miseq	300966	907	0.30	8.12	701590	9794710	Goat
RV2489	17/01/09	2*Miseq	2665913	3359	0.13	16.43	699900	9794413	Domestic dog
RV2490	07/08/08	Miseq	1013605	8526	0.84	79.09	702122	9796402	Domestic dog
RV2491	02/08/08	Miseq	1421454	10733	0.76	101.98	696748	9794246	Domestic dog
RV2492	03/11/07	Miseq	1642925	12895	0.78	10**0.9**3	685837	9799418	Domestic dog
RV2493	31/07/08	Miseq	1304543	4780	0.37	44.57	700689	9793979	Domestic dog
RV2495	12/07/08	Miseq	1946750	14768	0.76	129.48	696385	9802796	Domestic dog
RV2497	22/06/07	Miseq/Nextseq	15591341	6031	0.04	63.98	678577	9807715	Domestic dog
RV2500	23/11/08	Miseq	557886	1536	0.28	15.86	698019	9804712	Domestic dog
RV2501	14/02/04	Miseq	1030023	11539	1.12	108.23	679372	9810508	Domestic dog
RV2503	15/02/09	Miseq	1325075	2774	0.21	28.64	686729	9761151	Wild cat
RV2767	12/09/09	Miseq	1039525	23560	2.27	249.79	665178	9777064	Domestic dog
RV2786	20/06/10	Miseq/Nextseq	15706254	10705	0.07	109.00	670876	9811501	Domestic dog

RV2788	05/08/10	Miseq	1952400	16416	0.84	181.53	668396	9794012	Domestic dog
RV2789	22/10/10	Miseq	825552	1753	0.21	16.94	686741	9795339	Domestic dog
RV2790	01/11/10	Miseq	190118	2054	1.08	20.87	678362	9807838	Domestic dog
RV2791	08/11/10	Miseq/Nextseq	15068608	20740	0.14	237.54	681329	9802616	Domestic dog
RV2792	07/01/11	Miseq/Nextseq	13995380	19366	0.14	215.26	670624	9824613	Jackal
RV2793	02/01/11	Miseq	149085	3900	2.62	38.81	657176	9822908	Cow
RV2794	12/01/11	Miseq	514797	2064	0.40	19.28	679391	9807221	Domestic dog
RV2795	16/01/11	Miseq	369231	12824	3.47	134.33	682879	9801080	Cow
RV2796	19/01/11	Miseq	1033943	56028	5.42	508.13	671651	9801501	Cow
RV2797	30/01/11	Miseq	1278027	25740	2.01	229.21	663782	9800181	Cow
RV2798	31/01/11	Miseq	1287235	69613	5.41	604.45	663782	9800181	Cow
RV2799	29/01/11	Miseq	1347238	4352	0.32	41.33	680129	9805797	Domestic dog
RV2800	17/02/11	Miseq	142582	2917	2.05	33.48	666237	9778079	Domestic dog
RV2858	21/02/11	Miseq	1503125	120336	8.01	935.41	646139	9824316	Goat
RV2859	01/03/11	Miseq	283796	8437	2.97	88.54	650148	9819597	Domestic dog
RV2861	11/05/11	Miseq	1359627	164027	12.06	1259.37	691423	9791388	Domestic dog
RV2862	11/05/11	Miseq	1177250	2121	0.18	21.18	674259	9805687	Domestic dog
RV2863	19/04/11	2*Miseq	448320	24262	5.41	194.02	672066	9805512	Domestic dog
RV2866	15/04/11	Miseq/Nextseq	17707815	1007	0.01	9.98	678461	9808213	Goat

RV2867	22/05/11	gadep	2965005	2967	0.10	18.65	677206	9802038	Domestic dog
RV2868	06/06/11	Miseq	703095	27276	3.88	272.91	678588	9807873	Cow
RV2870	01/06/11	Miseq	529286	19381	3.66	154.52	697216	9792726	Domestic dog
RV2871	17/06/11	Amplicon seq: 454	46796	8850	18.91	25.34	701315	9794641	Domestic dog
RV2873	20/06/11	Miseq	2235772	35378	1.58	359.16	687042	9796792	Domestic dog
RV2875	29/06/11	Miseq	407187	9927	2.44	99.27	680196	9811835	Domestic dog
RV2877	30/06/11	Miseq	4169008	168447	4.04	1494.91	679837	9808655	Goat
RV2878	03/07/11	Miseq	146105	1933	1.32	21.85	689579	9796947	Domestic dog
RV2879	10/06/11	Miseq	696671	6806	**0.9**8	68.17	698379	9806085	Domestic dog
RV2880	02/07/11	Miseq	795216	8731	1.10	92.11	678588	9807873	Cow
RV2881	07/07/11	Miseq	1101481	12573	1.14	117.68	698357	9806963	Cow
RV2882	23/07/11	gadep	2896363	35151	1.21	88.58	676741	9807751	Domestic dog
RV2883	23/07/11	Miseq/Nextseq	20693938	21557	0.10	221.41	686525	9797308	Domestic dog
RV2884	21/07/11	Miseq	1107607	9926	**0.9**0	102.65	675366	9806739	Cow
RV2885	24/07/11	Miseq	969814	3303	0.34	33.06	694855	9798730	Goat
RV2886	21/07/11	Amplicon seq: 454	48056	16226	33.76	64.02	701234	9794942	Goat
RV2887	25/07/11	Miseq	1044324	2444	0.23	24.46	695864	9798032	Domestic dog
RV2888	29/07/11	Miseq	1517717	19413	1.28	189.29	700262	9793490	Domestic dog

RV2889	03/08/11	Miseq	1395952	5503	0.39	50.33	693407	9791698	Domestic dog
RV2890	03/08/11	Miseq	1725539	16406	**0.9**5	155.06	651666	9819798	Cow
RV2891	10/08/11	Miseq	1290933	54121	4.19	414.53	680536	9813026	Civet
RV2892	12/08/11	gadep	2796394	50188	1.79	91.09	676741	9807751	Domestic dog
RV2893	01/08/11	Miseq	1637158	31071	1.90	278.08	651666	9819798	Cow
RV2894	15/08/11	Miseq	2125096	188049	8.85	1469.55	653024	9822353	Domestic dog
RV2895	18/08/11	Miseq	2273420	19720	0.87	210.02	678334	9797356	Domestic dog
RV2896	19/08/11	Miseq	1104977	7214	0.65	68.67	681869	9798035	Domestic dog
RV2897	02/09/11	Miseq	984501	9581	**0.9**7	92.68	678907	9808303	Domestic dog
RV2898	18/09/11	Miseq/Nextseq	19197193	20664	0.11	206.05	658269	9804143	Domestic dog
RV2899	23/09/11	Miseq	897751	8818	**0.9**8	85.53	658245	9804462	Domestic dog
RV2900	27/09/11	Miseq	1326869	8924	0.67	84.32	656396	9803751	Domestic dog
RV2901	22/09/11	Miseq	522728	10710	2.05	106.87	653179	9802910	Domestic dog
RV2902	24/09/11	2*Miseq	1204863	5234	0.43	43.68	684532	9790009	Domestic dog
RV2903	05/10/11	Miseq	1219764	6605	0.54	62.45	657232	9804521	Domestic dog
RV2906	08/09/11	Miseq	1283722	19254	1.50	170.85	681338	9808012	Cat
RV2907	16/10/11	Miseq	1353429	7867	0.58	71.47	700560	9803728	Cow
RV2909	12/11/11	Miseq	893349	20317	2.27	194.08	663440	9804477	Cow

RV3047*	10/08/12	Nextseq	14822146	193963	1.31	1634.14	655560	9809040	Domestic dog
RV3048*	17/09/12	Nextseq	13774437	92882	0.67	865.92	690718	9803251	Sheep
RV3049*	04/10/12	Nextseq	14925758	68578	0.46	634.13	689087	9797146	Domestic dog
RV3050*	02/09/12	Nextseq	12716032	689856	5.43	5022.73	671593	9804068	Domestic dog
RV3051*	12/10/12	Nextseq	13274035	33398	0.25	308.65	684769	9779316	Domestic dog
RV3052*	13/11/12	Nextseq	10917330	110185	1.01	1077.20	643810	9813739	Domestic dog
RV3053*	10/11/12	Nextseq	11787802	61379	0.52	600.13	646262	9815621	Domestic dog
RV3057*	08/12/12	Nextseq	13502719	629341	4.66	4329.86	655999	9810468	Domestic dog
RV3058*	18/01/13	Nextseq	11873546	143723	1.21	1031.42	648713	9819325	Domestic dog
RV3059*	11/01/13	Nextseq	12420277	786476	6.33	4587.68	654132	9825734	Domestic dog
RV3060*	04/01/13	Nextseq	12150398	316716	2.61	2450.02	654509	9823226	Domestic dog
RV3061*	08/01/13	Nextseq	12380235	386682	3.12	2883.85	651604	9821249	Domestic dog
RV3063*	08/01/13	Nextseq	11500684	37364	0.32	365.20	691639	9810370	Domestic dog
RV3064*	02/01/13	Nextseq	11829231	48612	0.41	469.26	687567	9796857	Domestic dog
RV3066*	23/02/13	Nextseq	10441152	159856	1.53	1487.65	652145	9807506	Cat
RV3067*	25/02/13	Nextseq	15872978	3697	0.02	35.46	653061	9777091	Domestic dog

RV3068*	11/04/13	Nextseq	11386195	304247	2.67	2459.48	696403	9791890	Goat
RV3070*	27/02/13	Nextseq	10573869	50029	0.47	504.04	673866	9807701	Goat
RV3071*	13/04/13	Nextseq	13361066	1493699	11.18	7981.02	680151	9814868	Domestic dog
RV3072*	10/04/13	Nextseq	10909432	1088443	9.98	6940.73	672356	9817934	Domestic dog
RV3073*	18/02/13	Nextseq	9431525	84544	**0.90**	816.23	697846	9808299	Domestic dog
RV3074*	01/02/13	Nextseq	13487972	780515	5.79	5369.35	658498	9825235	Domestic dog
RV3075*	31/01/13	Nextseq	12309563	85376	0.69	834.07	658498	9825235	Domestic dog
RV3078*	26/05/13	Nextseq	13267270	324339	2.44	2550.76	684594	9779133	Cow
RV3079*	18/07/13	Nextseq	13140710	10369	0.08	103.78	700214	9793780	Domestic dog
RV3080*	17/07/13	Nextseq	12398079	71735	0.58	581.75	700300	9793350	Domestic dog
RV3082*	03/09/13	Nextseq	13023573	46367	0.36	485.26	670878	9795201	Domestic dog
RV3084*	22/08/13	Nextseq	9206525	18253	0.20	179.18	697865	9803607	Domestic dog
RV3085*	06/09/13	Nextseq	11789633	64210	0.54	590.80	683848	9810611	Domestic dog
RV3086*	06/09/13	Nextseq	13831121	298860	2.16	2439.13	697902	9792902	Cow
RV3087*	06/09/13	Nextseq	17902151	12527	0.07	126.18	699819	9795099	Cow
RV3088*	25/04/13	Nextseq	18221434	358845	1.97	2936.54	691229	9812934	Cow
RV3090	25/11/11	Miseq	129980	10825	8.33	104.41	696825	9790124	Cow
RV3091	05/12/11	Miseq/Nextseq	18376860	50860	0.28	496.86	696698	9792023	Domestic dog

RV3092	16/12/11	Miseq	1045220	2104	0.20	18.69	680305	9798935	Domestic dog
RV3093	22/12/11	Miseq	495576	4371	0.88	42.70	681720	9798404	Domestic dog
RV3094	21/12/11	Miseq	1100604	3942	0.36	35.79	699127	9791922	Domestic dog
RV3096	30/01/12	Miseq	758254	2900	0.38	24.44	658200	9804251	Domestic dog
RV3097	19/01/12	Miseq	384498	12069	3.14	113.75	658323	9804533	Domestic dog
RV3098	26/01/12	Miseq	277003	7448	2.69	73.18	658761	9804127	Domestic dog
RV3099	19/01/12	Miseq	1440888	30135	2.09	233.29	696747	9790182	Cow
RV3100	19/02/12	Miseq	668576	2648	0.40	24.11	698140	9804256	Domestic dog
RV3101	06/03/12	Miseq	416501	10348	2.48	86.78	695255	9803449	Jackal
RV3102	08/03/12	Miseq	291477	13309	4.57	123.32	700597	9803335	Goat
RV3103	10/12/11	Miseq/Nextseq	17549986	11210	0.06	109.93	653049	9821771	Domestic dog
RV3104	22/12/11	Miseq	1145840	11000	**0.9**6	95.11	648913	9823033	sheep
RV3105	22/03/12	Miseq	1009153	1365	0.14	13.11	661920	9803801	Domestic dog
RV3107	09/04/12	Miseq	959083	31920	3.33	292.64	656704	9803322	Domestic dog
RV3109	12/04/12	Miseq	409667	24684	6.03	259.04	656691	9803564	Domestic dog
RV3110	30/04/12	Miseq	346269	6180	1.78	65.54	686063	9798923	Domestic dog
RV3111	29/04/12	Miseq	897329	2674	0.30	26.85	650984	9807685	Domestic dog

RV3113	06/05/12	Miseq/Nextseq	18115513	68797	0.38	761.18	685149	9792574	Domestic dog
RV3114	06/05/12	Miseq	854689	16344	1.91	159.71	685205	9792549	Jackal
RV3115	11/05/12	Miseq	800250	3034	0.38	29.22	680871	9782055	Domestic dog
RV3117	12/05/12	Miseq	2331749	6102	0.26	57.60	669765	9795977	Domestic dog
RV3119	17/05/12	Miseq/Nextseq	15519701	7634	0.05	81.87	680871	9782055	Domestic dog
RV3121	28/05/12	Miseq	1321266	14930	1.13	132.03	655915	9809872	Domestic dog
RV3122	14/05/12	Miseq	1328914	5972	0.45	56.12	662827	9818718	Domestic dog
RV3123	26/04/12	Miseq	1530487	7221	0.47	65.08	647442	9799177	Domestic dog
RV3124	07/05/12	Miseq	903958	9073	1.00	86.74	657028	9804565	Domestic dog
RV3125	06/06/12	Miseq/Nextseq	16181484	10504	0.06	105.89	658003	9809433	Goat
RV3127	11/06/12	Miseq	461052	20729	4.50	204.27	681576	9797817	Domestic dog
RV3128	16/06/12	Miseq	2044844	9514	0.47	88.78	674458	9797076	Donkey
RV3129	25/06/12	Miseq	871812	41446	4.75	408.32	662827	9819718	Sheep
RV3130	28/06/12	Miseq	155506	10100	6.49	103.35	662827	9818718	Goat
RV3131*	07/07/12	Miseq	1518274	12181	0.80	113.24	685498	9797037	Domestic dog
RV3132	02/04/12	Miseq	489623	1478	0.30	13.93	656028	9801566	Domestic dog
RV3133	04/07/12	Miseq/Nextseq	14589139	11715	0.08	118.01	700856	9800221	Domestic dog

RV3134*	19/07/12	Miseq	1280905	3509	0.27	34.25	689067	9791597	Domestic dog
RV3135*	19/07/12	Miseq	1327972	10241	0.77	95.70	695591	9788282	Domestic dog
RV3136*	16/07/12	Miseq	993550	15864	1.60	153.90	678119	9799572	Cow
RV3137*	21/07/12	Miseq	673235	14221	2.11	140.50	682623	9783025	Goat
RV3138*	16/07/12	Miseq	1634873	13081	0.80	126.10	677206	9805047	Domestic dog
RV3139*	07/07/12	Miseq	1500822	7764	0.52	74.40	685498	9797937	Domestic dog
RV3140*	27/07/12	Miseq	1907863	99500	5.22	737.59	656945	9804374	Domestic dog
RV3145*	08/08/12	Miseq/Nextseq	15216797	10232	0.07	110.25	658701	9825071	Hyena
RV3146	26/05/12	Miseq/Nextseq	16819682	13702	0.08	135.45	697865	9803637	Civet
RV3149	23/01/12	Miseq	365335	802	0.22	6.23	657742	9809576	Domestic dog
RV3150	24/03/12	Miseq	1194163	43373	3.63	407.75	682197	9791269	Domestic dog
RV3151	07/01/12	Miseq	105666	6227	5.89	60.67	648913	9823033	Cow
RV3152	31/05/12	Miseq/Nextseq	14910632	4003	0.03	42.77	695002	9803207	NA
RV3153	23/01/12	Nextseq	12054558	62937	0.52	637.73	694625	9806377	Goat
RV3154	29/04/12	Nextseq	14618752	401835	2.75	3228.77	658670	9802617	Domestic dog

Table C.2: Pearson correlations between cost surfaces representing the effect of different landscape predictors on rabies virus diffusion. Predictor combinations are indicated in the first column with the following abbreviations: dd=dog density, dem=elevation, hdr=human to dog ratio, ibd=isolation by distance, susc=susceptibles, vacc=% vaccination coverage.

Predictors	Correlation
hdr,vacc	0.929471451
hdr,dd	0.938373837
hdr,slope	-0.107690061
hdr,road	0.032313333
hdr,dem	-0.084158885
hdr,river	0.039937272
hdr,susc	0.939061673
vacc,dd	0.973992055
vacc,slope	-0.070345425
vacc,road	0.03174924
vacc,dem	-0.089611481
vacc,river	0.041504086
vacc,susc	0.97639971
dd,slope	-0.068338848
dd,road	0.041296148
dd,dem	-0.137977376
dd,river	0.044736621
dd,susc	0.999786683
slope,road	-0.022397718
slope,dem	-0.067054847
slope,river	0.007311466
slope,susc	-0.069512983
road,dem	0.001631024
road,river	0.005224623
road,susc	0.041322333
dem,river	-0.066428911
dem,susc	-0.133284777
river,susc	0.044550173

Table C.3: Pearson correlations between landscape predictor resistance distances at different levels of spatial discretisation (k=number of discrete clusters) tested in GLM models in Chapter 4. Predictor combinations are indicated in the first column with abbreviations as in Table C.3. Correlations greater than or equal to 0.9 are highlighted in bold.

Predictors	k5	k6	k7	k8	k9	k10	k11	k12	k13	k14	k15
dd,dem	0.47	0.65	0.35	0.44	0.38	0.43	0.28	0.47	0.42	0.36	0.37
dd,slope	0.78	0.61	0.22	0.54	0.21	0.41	0.27	0.05	0.35	0.21	-0.13
dd,vacc	0.30	0.55	0.23	0.31	0.36	-0.22	0.28	0.20	0.09	0.40	0.38
dd,road	0.22	0.69	0.69	0.26	0.37	0.59	0.47	0.52	0.69	0.51	0.51
dd,river	0.14	0.34	0.33	0.41	0.39	0.23	0.36	0.27	0.55	0.42	0.29
dd,hdr	-0.01	0.19	-0.10	-0.04	0.13	-0.32	-0.14	-0.04	0.01	0.07	0.14
dd,susc	**0.97**	**0.99**	**0.98**	**0.98**	**0.98**	**0.96**	**0.98**	**0.96**	**0.96**	**0.97**	**0.98**
dem,slope	0.28	0.41	0.15	0.47	0.33	0.30	0.42	0.14	0.38	0.36	0.00
dem,vacc	0.49	0.55	0.74	0.09	0.17	0.37	0.31	0.48	0.34	0.39	0.52
dem,road	0.52	0.52	0.46	0.49	0.57	0.41	0.53	0.54	0.32	0.49	0.45
dem,river	0.26	0.78	0.75	0.75	0.79	0.21	0.76	0.71	0.71	0.82	0.60
dem,hdr	0.42	0.47	0.58	0.27	0.39	0.32	0.35	0.53	0.45	0.42	0.58
dem,susc	0.51	0.69	0.42	0.52	0.43	0.46	0.33	0.55	0.45	0.41	0.43
slope,vacc	-0.16	-0.16	-0.40	-0.13	-0.36	-0.23	-0.07	-0.34	-0.41	-0.11	-0.22
slope,road	-0.32	0.06	0.04	-0.13	0.00	0.05	0.09	-0.07	0.06	0.03	0.03
slope,river	0.16	0.51	0.08	0.52	0.35	0.32	0.43	0.25	0.43	0.43	0.41
slope,hdr	-0.49	-0.44	-0.47	-0.46	-0.55	-0.51	-0.45	-0.38	-0.44	-0.43	-0.54
slope,susc	0.75	0.57	0.08	0.51	0.11	0.32	0.25	-0.09	0.22	0.15	-0.17
vacc,road	0.40	0.54	0.43	0.02	0.06	0.10	-0.03	0.19	0.17	0.06	0.15
vacc,river	0.43	0.10	0.44	-0.13	0.02	-0.13	0.18	0.13	-0.03	0.22	0.23
vacc,hdr	**0.91**	0.76	0.72	0.73	0.80	**0.90**	0.65	0.79	0.89	0.71	0.77
vacc,susc	0.48	0.64	0.36	0.35	0.44	-0.08	0.32	0.37	0.24	0.49	0.45
road,river	-0.29	0.06	0.14	0.09	0.22	0.15	0.31	0.33	0.32	0.34	0.26
road,hdr	0.51	0.21	0.00	0.15	0.18	-0.03	-0.08	0.24	0.04	-0.07	0.04
road,susc	0.14	0.70	0.75	0.33	0.45	0.63	0.57	0.63	0.68	0.60	0.58
river,hdr	0.42	0.25	0.57	0.00	0.22	-0.11	0.15	0.30	0.11	0.23	0.14
river,susc	0.28	0.34	0.38	0.42	0.39	0.32	0.41	0.28	0.46	0.43	0.32
hdr,susc	0.14	0.23	0.08	0.08	0.28	-0.16	-0.10	0.17	0.14	0.15	0.23

APPENDIX D

Chapter 5 Appendix

D.1 Inference of the transmission tree based on spatio-temporal data, pathogen genetic data, and contact tracing data

Ideally, knowing who infected whom for every transmissions would allows to finely determine the risk factors. However, contact tracing is only partial (and contain a part of uncertainty about the effective source; i.e. a contact does not automatically implies a transmission).

Alternatively, modeling and statistical approaches have been developed to infer who-infected-whom based on spatio-temporal and genomic data (Hall *et al.*, 2015; Jombart *et al.*, 2014; Mollentze *et al.*, 2014b; Morelli *et al.*, 2012; Ypma *et al.*, 2012, 2013). One of the approaches recently proposed is based on an extension of stochastic Susceptible-Exposed-Infectious-Removed (SEIR) models and was applied to infer rabies transmissions in Kwa Zulu Natal province in South Africa (Mollentze *et al.*, 2014b). This approach combines heterogeneous and multi-scale processes and data: it links the epidemiological scale and the micro-evolutionary scale.

The approach of Mollentze *et al.* (2014b) is based on a genetic-space-time model, which combines (i) an individual-based, spatial, semi-Markov SEIR model for the spatio-temporal dynamics of the pathogen, and (ii) a Markovian evolutionary model for the temporal evolution of genetic sequences of the pathogen. The resulting model is a state-space model including latent vectors of high dimension (e.g. the transmission tree, the infection times, the unobserved sequences of the pathogen at the transmission times). Mollentze *et al.* (2014b) estimated model parameters and latent variables in the Bayesian framework with an approximate MCMC algorithm. Soubeyrand (2014) proposed an other approximate MCMC algorithm, which was shown to improve the inference about the transmission tree based on a simulation study. Moreover, Soubeyrand (2014) shown how to handle incompleteness of genetic data, i.e. the missing pathogen sequences (pathogen sequences are not observed for all the hosts in the epidemiological data base).

In the case of rabies in Tanzania, we applied the approach presented in Soubeyrand (2014) and extended it by incorporating contact tracing data and a zero-inflated dispersal kernel. Additionally, we made a distinction like Morelli *et al.* (2012) did between the observation time, at which the host is observed as infected and sequenced, and the end time, at which the host is removed. In the following, we do not detail the whole approach but only highlight the extensions mentioned above.

D.1.1 Posterior distribution

We consider the joint posterior distribution $p(J, T^{inf}, L, D, \theta \mid data)$ of the transmission tree J, infection times $T^{inf} = (T_1^{inf}, \ldots, T_n^{inf})$, exposed (or latency) durations $L = (L_1, \ldots, L_n)$, infectious durations $D = (D_1, \ldots, D_n)$ before observations, and parameters θ that contains infection and dispersal parameters $\alpha = (\alpha_0, \alpha_1, \alpha_2) = (\alpha_0, \alpha_1, (\alpha_{2,1}, \alpha_{2,2}, \alpha_{2,3}))$, latency parameters $\beta = (\beta_1, \beta_2)$, infectiousness parameters $\delta = (\delta_1, \delta_2)$, mutation parameters $\mu = (\mu_1, \mu_2, \mu_3)$ and the date t_{exo} of the exogenous sequence S_{exo}. The exogenous sequence and its date are used to model the infections from unobserved hosts or, in other words, the infections from the disease reservoir. The use of an exogenous sequence and its date allows us to easily handle the incompleteness of epidemiological data, i.e. the missing infecting hosts.

The transmission tree J is a function from $\{1,\dots,n\}$ to $\{0,1,\dots,n\}$ that states who infected whom: an observed individual i is infected by a pathogen source $j = J(i)$ that is either another observed individual $j \in \{1,\dots,n\}$, $j \neq i$, or the disease reservoir (exogenous source) denoted by 0.

Data are observation times $T^{obs} = (T_1^{obs},\dots,T_n^{obs})$, removal times $T^{end} = (T_1^{end},\dots,T_n^{end})$, central locations $X = (x_1,\dots,x_n)$ of observed individuals, their abilities to spread the disease $A = (A_1,\dots,A_n)$, observed sequences $S^{obs} = \{S_1(T_1^{obs}),\dots,S_n(T_n^{obs})\}$ at the observation times, and contact tracing information $\mathcal{C}$. The posterior distribution is:

$$
\begin{aligned}
p(J,T^{inf},L,D,\theta \mid data) &= p(J,T^{inf},L,D,\theta \mid S^{obs},T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C}) \\
&\propto p(S^{obs} \mid J,T^{inf},L,D,\theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C})p(J,T^{inf},L,D,\theta \mid T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C}) \\
&= p(S^{obs} \mid J,T^{inf},L,D,\theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C})p(J,T^{inf} \mid L,D,\theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C}) \\
&\quad \times p(L,D \mid \theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C})p(\theta \mid X,\mathcal{C}) \\
&= p(S^{obs} \mid J,T^{inf},L,D,\theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C})p(J,T^{inf} \mid L,D,\theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C}) \\
&\quad \times p(L,D \mid \theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C})p(\alpha_0,\alpha_1,\beta,\delta,\mu,t_{exo})p(\alpha_2 \mid X,\mathcal{C}) \\
&\propto p(S^{obs} \mid J,T^{inf},L,D,\theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C})p(J,T^{inf} \mid L,D,\theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C}) \\
&\quad \times p(L,D \mid \theta,T^{obs},T^{end},X,S_{\text{exo}},A,\mathcal{C})p(\alpha_0,\alpha_1,\beta,\delta,\mu,t_{exo})p(\mathcal{C} \mid \alpha_2,X)p(\alpha_2)
\end{aligned}
\tag{D.1}
$$

where $\propto$ means "proportional to" (the multiplicative constant does not depend on the unknowns (J,T^{inf},L,D,θ)), $p(S^{obs} \mid J,T^{inf},L,D,\theta,T^{end},X,S_{\text{exo}})$ is called the genetic likelihood, $p(J,T^{inf} \mid L,D,\theta,T^{end},X,S_{\text{exo}})$ is called the transmission likelihood, $p(L,D \mid \theta,T^{end},X,S_{\text{exo}})$ is the distribution of latency and infectious durations, $p(\mathcal{C} \mid \alpha_2,X)$ is the contact likelihood, and $p(\alpha_0,\alpha_1,\beta,\delta,\mu,t_{exo})p(\alpha_2)$ is the prior distribution of parameters which is supposed to not depend on explanatory variables.

We refer to Soubeyrand (2014) for the expression of the genetic likelihood and its approximation.

The distribution of latency durations and infectious durations before observations is assumed to be the following product of gamma probability densities:

$$
\begin{aligned}
p(L,D \mid \theta,T^{obs},T^{end},X,S_{\text{exo}}) &= p(L,D \mid \theta) \\
&= \prod_{i=1}^{I} \gamma(L_i;\beta_1,\beta_2)\gamma(D_i;\delta_1,\delta_2),
\end{aligned}
$$

where $\gamma(\cdot;a,b)$ is the probability distribution function of the gamma distribution parameterized by (a,b).

The transmission likelihood, the contact likelihood and the prior distribution of parameters are specified in the following subsections.

D.1.2 Transmission likelihood

The transmission likelihood $p(J, T^{inf} \mid L, D, \theta, T^{obs}, T^{end}, X, S_{\text{exo}}, A, \mathcal{C})$ can be written:

$$\begin{aligned} p(J, T^{inf} \mid L, D, \theta, T^{obs}, T^{end}, X, S_{\text{exo}}, A, \mathcal{C}) &= p\left(J(1), T_1^{inf} \mid L, D, \theta, T^{obs}, T^{end}, X, A, \mathcal{C}\right) \\ &\times \prod_{i=2}^{I} p\left(J(i), T_i^{inf} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{inf}, L, D, \theta, T^{obs}, T^{end}, X, A, \mathcal{C}\right), \end{aligned} \tag{D.2}$$

where the index i is sorted with respect to increasing infection times T_i^{inf}, $J\{1:(i-1)\} = (J(1), \ldots, J(i-1))$ for $i > 1$, $T_{1:(i-1)}^{inf} = (T_1^{inf}, \ldots, T_{i-1}^{inf})$ for $i > 1$, and by assuming that the transmission dynamics does not depend on the exogenous sequence S_{exo}.

In the following, contact data is assumed to be the set of pairs of infected hosts that were observed to be in contact:

$$\mathcal{C} = \{(i,j) \in \{1, \ldots, I-1\} \times \{i+1, \ldots, I\} : i \text{ and } j \text{ were in contact}\}.$$

Moreover, contact data for host i is the set of hosts that were observed to be in contact with i:

$$\mathcal{C}_i = \{j \in \{1, \ldots, I\} - \{i\} : i \text{ and } j \text{ were in contact}\}.$$

The abilities of hosts to spread the disease $A = (A_1, \ldots, A_n)$ are supposed to be binary variables depending on which species the host belong to: A_i is equal to 1 if host i belongs to a species able to contaminate susceptibles (dogs, jackals, wild cats...), and 0 otherwise (livestock).

Concerning the first term of the right-hand side of Equation (D.2), each host has the same chance $(1/I)$ to be infected first (by an external source $J(1) = 0$), and its infection time is assumed to be less than or equal to the first observation time ($\min\{T^{end}\}$):

$$p\left(J(1), T_1^{inf} \mid L, D, \theta, T^{obs}, T^{end}, X, A, \mathcal{C}\right) = \frac{1}{I} \times \mathbf{1}(T_1^{inf} \leq \min\{T^{end}\}),$$

where $\mathbf{1}$ is the indicator function.

Subsequent infections (i.e. for $i > 1$) occur with the following probabilities:

$$\begin{aligned} & p\left(J(i), T_i^{inf} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{inf}, L, D, \theta, T^{end}, X, \mathcal{C}\right) \\ &= p\left(J(i), T_i^{inf} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{inf}, L, \theta, T^{end}, X, \mathcal{C}_i\right) \\ &= \exp\left(-\alpha_0 (T_i^{inf} - T_1^{inf}) - \int_{T_1^{inf}}^{T_i^{inf}} \sum_{\substack{j=1 \\ j \neq i}}^{I} \alpha_1 \mathbf{1}(T_j^{inf} + L_j \leq t \leq T_j^{end}) A_j w(x_j - x_i) dt \right. \\ &\qquad \left. - \sum_{\substack{j \in \mathcal{C}_i \\ j \neq J(i)}} \epsilon \rho \alpha_1 \mathbf{1}(T_j^{end} \leq T_i^{inf}) A_j w(0) \right) \\ &\quad \times \Big(\alpha_0 \mathbf{1}\{J(i) = 0\} \\ &\qquad + \alpha_1 \mathbf{1}(T_{J(i)}^{inf} + L_{J(i)} \leq T_i^{inf} \leq T_{J(i)}^{end}) A_j w(x_{J(i)} - x_i) \mathbf{1}\{J(i) \neq 0 \text{ and } J(i) \notin \mathcal{C}_i\} \\ &\qquad + \rho \alpha_1 \mathbf{1}(T_{J(i)}^{inf} + L_{J(i)} \leq T_i^{inf} \leq T_{J(i)}^{end}) A_j w(0) \mathbf{1}\{J(i) \neq 0 \text{ and } J(i) \in \mathcal{C}_i\} \Big) \end{aligned} \tag{D.3}$$

where the exponential term is the probability that host i has not been infected between times T_1^{inf} and T_i^{inf}, and the second term is the probability density that host i has been infected by $J(i)$ at time T_i^{inf}. Here, if $J(i) > 0$ the source is observed, while the source is external to the dataset (an introduction) if $J(i) = 0$. α_0 is the infection strength of the exogenous sources, assumed to be constant in time and space, α_1 is the infection strength of an observed source, and w is a parametric dispersal kernel. This kernel is assumed to be a zero-inflated power-exponential kernel parametrized by $\alpha_2 = (\alpha_{2,1}, \alpha_{2,2}, \alpha_{2,3}) \in \mathbb{R}_+^* \times \mathbb{R}_+^* \times [0,1]$ and satisfying, for all $x \in \mathbb{R}^2$:

$$w(x) = \alpha_{2,3} + (1-\alpha_{2,3}) \frac{\alpha_{2,2}}{2\pi(\alpha_{2,1})^2 \Gamma\left(\frac{2}{\alpha_{2,2}}\right)} \exp\left\{-\left(\frac{||x||}{\alpha_{2,1}}\right)^{\alpha_{2,2}}\right\}. \tag{D.4}$$

D.1.3 Details of Equation (D.3)

When a contact is observed, the contact does not occur during the whole study period but is rather short in time. This point was taken into account to obtain Equation (D.3) and was formalized by considering the following time-varying contact information:

$$\mathcal{C}_i(t) = \{j \in \{1, \ldots, I\} - \{i\} : i \text{ and } j \text{ were in contact at time } t\}.$$

It has to be noted that we do not have this information at our disposal but we only know $\mathcal{C}_i$.

To take into account contact tracing, we further introduce a time-varying dispersal kernel $\tilde{w}$ which is specified in the next equation. Infection of i depends on the following time-varying rate of transmission:

$$\begin{aligned}\lambda_i(t) &= \alpha_0 + \sum_{\substack{j=1 \\ j\neq i}}^{I} \alpha_1 \mathbf{I}_j(t) A_j \tilde{w}(x_j - x_i; t) \\ &= \alpha_0 + \sum_{\substack{j=1 \\ j\neq i}}^{I} \alpha_1 \mathbf{I}_j(t) A_j \{w(x_j - x_i)\mathbf{1}(j \notin \mathcal{C}_i(t)) + \rho w(0)\mathbf{1}(j \in \mathcal{C}_i(t))\},\end{aligned}$$

where $\mathbf{I}_j(t) = \mathbf{1}(T_j^{inf} + L_j \leq t \leq T_j^{end})$. When i and j are in contact, the distance is reduced to zero and the multiplicative factor ρ modifies the risk of transmission (a priori, ρ should be greater than 1).

The computation of $P_i = p\left(J(i), T_i^{inf} \mid J\{1:(i-1)\}, T_{1:(i-1)}^{inf}, L, D, \theta, T^{obs}, T^{end}, X, A, \mathcal{C}\right)$ is made as follows.

In P_i, the conditioning by (L, D, T^{obs}) implies that T_i^{inf} has to be equal to $T_i^{obs} - D_i - L_i$. Therefore, P_i reduces to the distribution of $J(i)$ and is the product of:

- $Q_i^{(1)}$, the probability that all $j \neq J(i)$ have not infected i until T_i^{inf}, and
- $Q_i^{(2)}$, the density probability that $J(i)$ infected i at time T_i^{inf}.

Therefore, for $j \neq i$, the time-varying rate of transmission from j to i is:

$$\lambda_{ij}(t) = \begin{cases} \alpha_0 & \text{if } j = 0 \\ \alpha_1 \mathbf{I}_j(t) A_j \{w(x_j - x_i)\mathbf{1}(j \notin \mathcal{C}_i(t)) + \rho w(0)\mathbf{1}(j \in \mathcal{C}_i(t))\} & \text{if } j \neq 0. \end{cases}$$

For $j \neq i$, the probability P_{ij} that j has not infected i until time T_i^{inf} is equal to:

$$P_{ij} = \exp\left(-\int_{T_1^{inf}}^{T_i^{inf}} \lambda_{ij}(t)dt\right),$$

which yields:

$$Q_i^{(1)} = \prod_{\substack{j=1 \\ j\neq i, j\neq J(i)}}^{I} P_{ij} = \prod_{\substack{j=1 \\ j\neq i, j\neq J(i)}}^{I} \exp\left(-\int_{T_1^{inf}}^{T_i^{inf}} \lambda_{ij}(t)dt\right).$$

The density probability $Q_i^{(2)}$ is equal to the temporal derivative assessed at time T_i^{inf} of the probability that $J(i)$ infected i at a time smaller than or equal to T_i^{inf}. Therefore,

$$\begin{aligned} Q_i^{(2)} &= \frac{\partial}{\partial t}(1 - P_{iJ(i)})\Big|_{t=T_i^{inf}} \\ &= \frac{\partial}{\partial t}\left\{1 - \exp\left(-\int_{T_1^{inf}}^{T_i^{inf}} \lambda_{iJ(i)}(t)dt\right)\right\}\Big|_{t=T_i^{inf}} \\ &= \lambda_{iJ(i)}(T_i^{inf}) \exp\left(-\int_{T_1^{inf}}^{T_i^{inf}} \lambda_{iJ(i)}(t)dt\right) \\ &= \lambda_{iJ(i)}(T_i^{inf}) P_{iJ(i)}. \end{aligned}$$

Let

- $P_i^{(1)} = Q_i^{(1)} P_{iJ(i)}$ be the probability that all j have not infected i until T_i^{inf}, and
- $P_i^{(1)} = \lambda_{iJ(i)}(T_i^{inf})$ the instantaneous rate of infection of i by $J(i)$ at time T_i^{inf},

such that $P_i = P_i^{(1)} P_i^{(2)}$. Then

$$\begin{aligned} P_i^{(1)} &= \exp\left(-\int_{T_1^{inf}}^{T_i^{inf}} \lambda_i(t)dt\right) \\ &= \exp\left(-\alpha_0(T_i^{inf} - T_1^{inf}) - \int_{T_1^{inf}}^{T_i^{inf}} \sum_{\substack{j=1 \\ j\neq i}}^{I} \alpha_1 \mathbf{I}_j(t) A_j w(x_j - x_i) \mathbf{1}(j \notin \mathcal{C}_i(t))dt \right. \\ &\qquad \left. - \int_{T_1^{inf}}^{T_i^{inf}} \sum_{\substack{j=1 \\ j\neq i}}^{I} \rho\alpha_1 \mathbf{I}_j(t) A_j w(0) \mathbf{1}(j \in \mathcal{C}_i(t))dt \right). \end{aligned} \tag{D.5}$$

Let us make the following assumptions:

(**A1**) Each contact occurs on a time interval $[\tau_{ij}, \tau_{ij} + \epsilon]$ (which depends on the hosts i and j in contact) such that ϵ is negligible with respect to the duration of the infectious periods;

(**A2**) If i and j were in contact (according to the contact tracing) and if $J(i) = j$, then the infection of i by j (at time T_i^{inf}) arised when the observed contact occurred, i.e. in the time interval $[\tau_{ij}, \tau_{ij} + \epsilon]$;

(**A3**) For all $j \in \mathcal{C}_i$ such that (a) $J(i) \neq j$ and (b) i was infected before the end of the removal time T_j^{end} of j, then the contact between i and j is supposed to have occurred before the infection of i and have been unsuccessful.

Using Assumption (**A1**), we make the following approximations for two terms occurring in the expression of $P_i^{(1)}$ in Equation (D.5): for all $j \neq i$,

$$\int_{T_1^{inf}}^{T_i^{inf}} \alpha_1 \mathbf{I}_j(t) A_j w(x_j - x_i) \mathbf{1}(j \notin \mathcal{C}_i(t)) dt \approx \int_{T_1^{inf}}^{T_i^{inf}} \alpha_1 \mathbf{I}_j(t) A_j w(x_j - x_i) dt, \tag{D.6}$$

and for all $j \in \mathcal{C}_i$,

$$\int_{T_1^{inf}}^{T_i^{inf}} \rho\alpha_1 \mathbf{I}_j(t) A_j w(0) \mathbf{1}(j \in \mathcal{C}_i(t)) dt \approx \epsilon\rho\alpha_1 \mathbf{I}_j(\tau_{ij}) A_j w(0) \mathbf{1}(j \in \mathcal{C}_i) \mathbf{1}(T_1^{inf} \leq \tau_{ij} < T_i^{inf}). \tag{D.7}$$

Under Assumptions (**A2**), if $j \in \mathcal{C}_i$, then the indicator function $\mathbf{1}(T_1^{inf} \leq \tau_{ij} < T_i^{inf})$ in Equation (D.7) is equal to one if and only if $J(i) = j$.

Under Assumptions (**A3**), if $j \in \mathcal{C}_i$ and $j \neq J(i)$, then $\mathbf{I}_j(\tau_{ij}) = \mathbf{1}(T_j^{end} \leq T_i^{inf})$.

Using the consequences of Assumptions (**A2**) and (**A3**), and using Equations (D.6-D.7), $P_i^{(1)}$ in Equation (D.5) can be approximated by:

$$\begin{aligned} P_i^{(1)} &\approx \exp\left(-\alpha_0(T_i^{inf} - T_1^{inf}) - \int_{T_1^{inf}}^{T_i^{inf}} \sum_{\substack{j=1 \\ j\neq i}}^{I} \alpha_1 \mathbf{I}_j(t) A_j w(x_j - x_i) dt \right. \\ &\qquad \left. - \sum_{\substack{j \in \mathcal{C}_i \\ j \neq J(i)}} \epsilon\rho\alpha_1 \mathbf{1}(T_j^{end} \leq T_i^{inf}) A_j w(0) \right) \\ &= \exp\left(-\alpha_0(T_i^{inf} - T_1^{inf}) - \int_{T_1^{inf}}^{T_i^{inf}} \sum_{\substack{j=1 \\ j\neq i}}^{I} \alpha_1 \mathbf{1}(T_j^{inf} + L_j \leq t \leq T_j^{end}) A_j w(x_j - x_i) dt \right. \\ &\qquad \left. - \sum_{\substack{j \in \mathcal{C}_i \\ j \neq J(i)}} \epsilon\rho\alpha_1 \mathbf{1}(T_j^{end} \leq T_i^{inf}) A_j w(0) \right). \end{aligned}$$

The product $\epsilon\rho$ is the ratio between the infection risk due to an host which is in the contact tracing during the contact and the infection risk cumulated over one time unit (i.e. a day) due to an host which is at a distance equal to zero.

Under Assumptions (**A2**),

$$P_i^{(2)} = \begin{cases} \alpha_0 & \text{if } J(i) = 0 \\ \alpha_1 \mathbf{1}(T_{J(i)}^{inf} + L_{J(i)} \leq T_i^{inf} \leq T_{J(i)}^{end}) A_j w(x_{J(i)} - x_i) & \text{if } J(i) \neq 0 \text{ and } J(i) \notin \mathcal{C}_i \\ \rho\alpha_1 \mathbf{1}(T_{J(i)}^{inf} + L_{J(i)} \leq T_i^{inf} \leq T_{J(i)}^{end}) A_j w(0) & \text{if } J(i) \neq 0 \text{ and } J(i) \in \mathcal{C}_i \end{cases}$$

Equation (D.3) is based on the equation of $P_i^{(1)}$ and $P_i^{(2)}$.

D.1.4 Contact likelihood

The contact likelihood satisfies:

$$p(\mathcal{C} \mid \alpha_2, X) = \prod_{(i,j)\in\mathcal{C}} w(x_i - x_j),$$

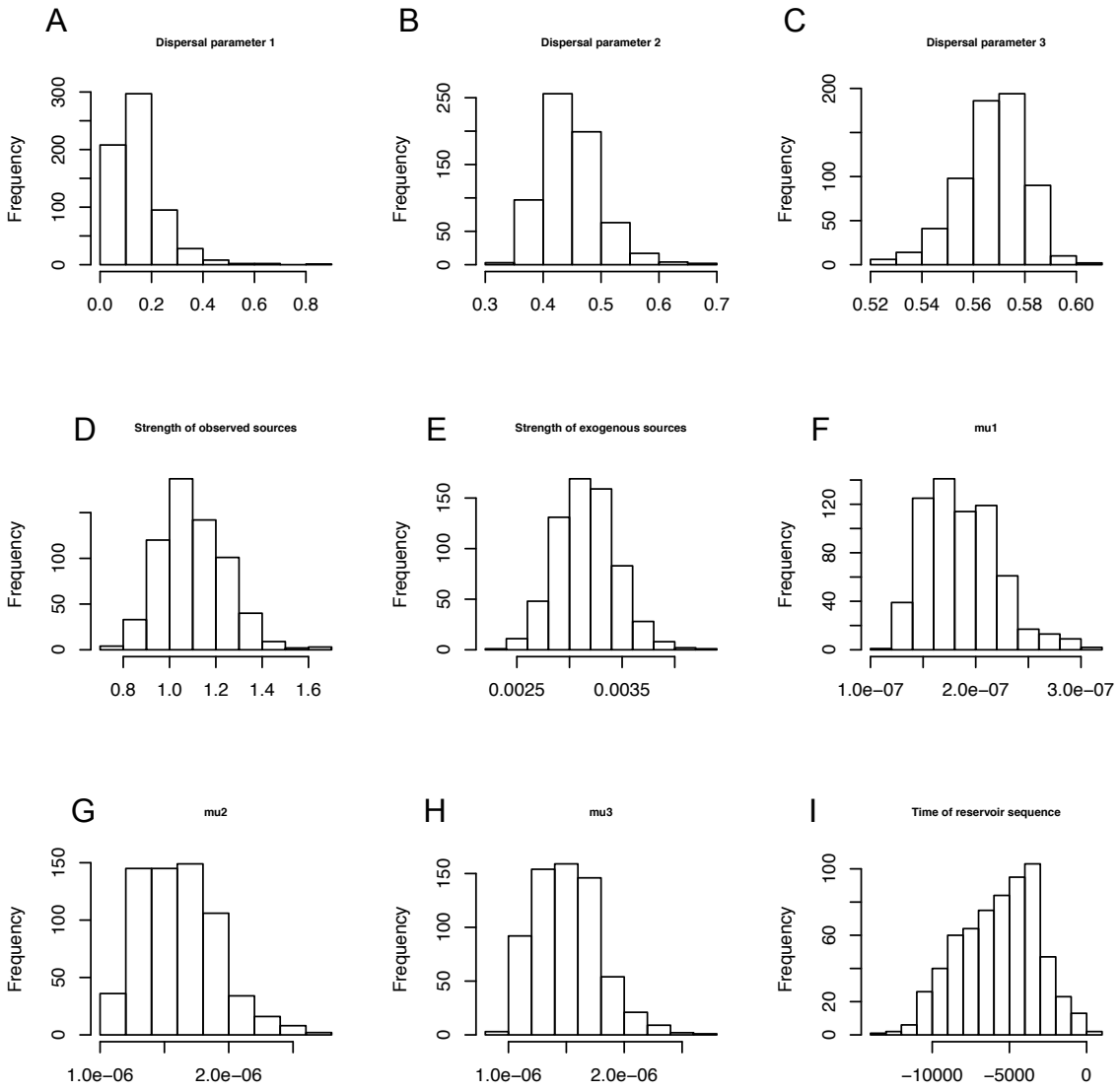


Figure D.1: Posterior distributions of parameters in Table 5.1 relating to dispersal (A-E); strength of observed (D) and exogeneous (E) sources; genetic mutation rates (G-H) and the estimated time of the reservoir sequence (I).

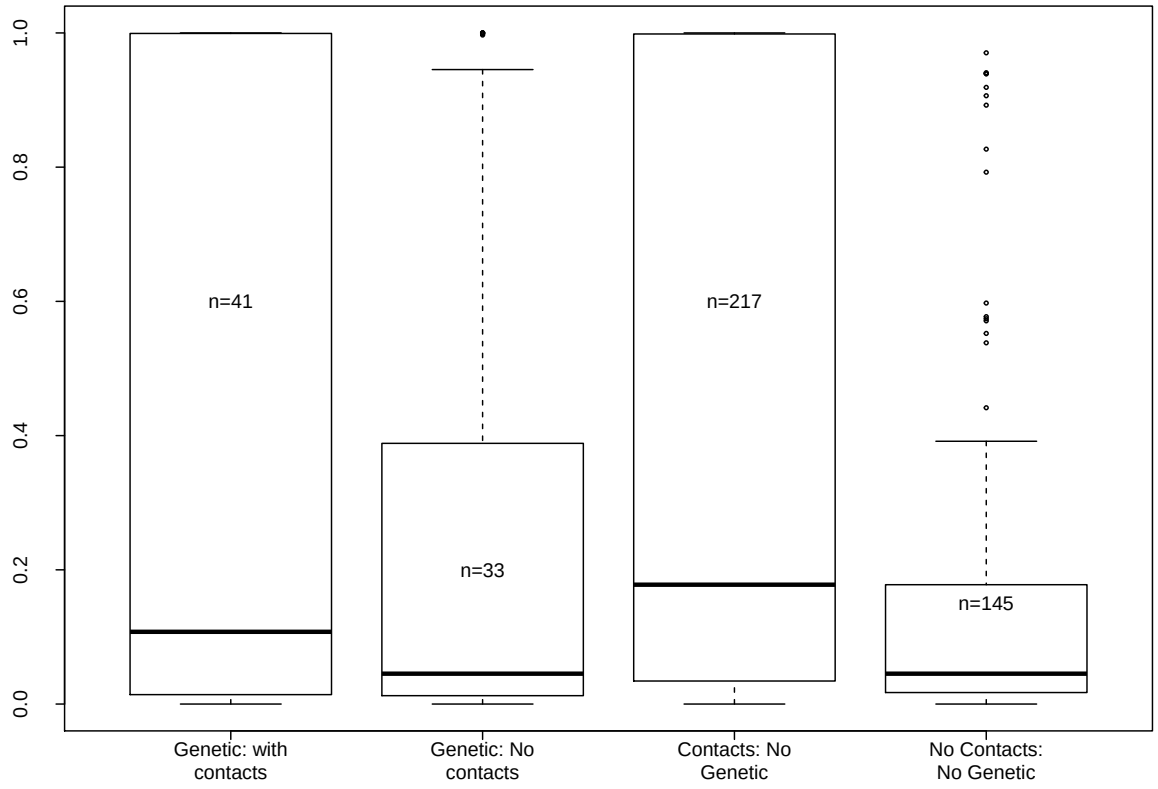


Figure D.2: Posterior probability of an observed source for cases with varying levels of observed data, showing (left to right) cases with observed genetic and contact tracing data; cases with observed genetic data but no contacts; cases with contact traced sources but no genetic data and cases with no genetic or contact data.

BIBLIOGRAPHY

Bibliography

Ahmed, K., Phommachanh, P., Vorachith, P., Matsumoto, T., Lamaningao, P., Mori, D., Takaki, M., Douangngeun, B., Khambounheuang, B., & Nishizono, A. (2015). Molecular Epidemiology of Rabies Viruses Circulating in Two Rabies Endemic Provinces of Laos, 2011âĂŞ2012: Regional Diversity in Southeast Asia. *PLOS Neglected Tropical Diseases*, **9**; page e0003,645. ISSN 1935-2735.

Anderson, C. D., Epperson, B. K., Fortin, M. J., Holderegger, R., James, P. M. A., Rosenberg, M. S., Scribner, K. T., & Spear, S. (2010). Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Molecular Ecology*, **19**(17); pages 3565–3575. ISSN 09621083.

Anderson, R. M. & May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*, volume 26. Oxford University Press. ISBN 0198545991.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data.

Archie, E. A., Luikart, G., & Ezenwa, V. O. (2009). Infecting epidemiology with genetics: a new frontier in disease ecology. *Trends in ecology & evolution*, **24**(1); pages 21–30. ISSN 0169-5347.

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut, A., & Suchard, M. a. (2012). BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic biology*, **61**(1); pages 170–3. ISSN 1076-836X.

Baddeley, A. & Turner, R. (2005). spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal Of Statistical Software*, **12**(6); pages 1–42. ISSN 15487660.

Ball, F. (1985). Front-wave velocity and fox habitat heterogeneity. In Bacon, P., editor, *Population Dynamics of Rabies in Wildlife*. Academic Press.

Banerjee, a. K. (1987). Transcription and replication of rhabdoviruses. *Microbiological reviews*, **51**(1); pages 66–87. ISSN 0146-0749.

Barton, H. D., Gregory, A. J., Davis, R., Hanlon, C. A., & Wisely, S. M. (2010). Contrasting landscape epidemiology of two sympatric rabies virus strains.

Bedford, T., Cobey, S., Beerli, P., & Pascual, M. (2010). Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS pathogens*, **6**(5); page e1000,918. ISSN 1553-7374.

Bedford, T., Suchard, M. a., Lemey, P., Dudas, G., Gregory, V., Hay, A. J., McCauley, J. W., Russell, C. a., Smith, D. J., & Rambaut, A. (2014). Integrating influenza antigenic dynamics with molecular evolution. *eLife*, **3**(3); pages 1–26. ISSN 2050-084X.

Beier, P., Majka, D. R., & Spencer, W. D. (2008). Forks in the road: Choices in procedures for designing wildland linkages. *Conservation Biology*, **22**(4); pages 836–851. ISSN 08888892.

Beier, P., Spencer, W., Baldwin, R. F., & Mcrae, B. H. (2011). Toward Best Practices for Developing Regional Connectivity Maps. *Conservation Biology*, **25**(5); pages 879–892. ISSN 08888892.

Beyer, H. L., Hampson, K., Lembo, T., Cleaveland, S., Kaare, M., & Haydon, D. T. (2011). Metapopulation dynamics of rabies and the efficacy of vaccination. *Proceedings. Biological sciences / The Royal Society*, **278**(1715); pages 2182–90. ISSN 1471-2954.

Bharti, N., Tatem, a. J., Ferrari, M. J., Grais, R. F., Djibo, A., & Grenfell, B. T. (2011). Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery. *Science*, **334**(6061); pages 1424–7. ISSN 1095-9203.

Biek, R., Henderson, J. C., Waller, L. A., Rupprecht, C. E., & Real, L. A. (2007). A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(19); pages 7993–8. ISSN 0027-8424.

Biek, R., Pybus, O. G., Lloyd-Smith, J. O., & Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution*, **30**(6); pages 306–313. ISSN 01695347.

Biek, R. & Real, L. a. (2010). The landscape genetics of infectious disease emergence and spread. *Molecular ecology*, **19**(17); pages 3515–31. ISSN 1365-294X.

Bielejec, F., Lemey, P., Baele, G., Rambaut, A., & Suchard, M. a. (2014). Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Systematic biology*, **63**(4); pages 493–504. ISSN 1076-836X.

Bielejec, F., Rambaut, A., Suchard, M. a., & Lemey, P. (2011). SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics (Oxford, England)*, **27**(20); pages 2910–2. ISSN 1367-4811.

Bingham, J., Foggin, C. M., Wandeler, a. I., & Hill, F. W. (1999). The epidemiology of rabies in Zimbabwe. 2. Rabies in jackals (Canis adustus and Canis mesomelas). *The Onderstepoort journal of veterinary research*, **66**(1); pages 11–23. ISSN 0030-2465.

Bivand, R., Pebesma, E., Gómez-Rubio, V., & NetLibrary, I. (2008). *Applied spatial data analysis with R*. Springer New York, Berlin Heidelberg NewYork HongKong London Milan Paris Tokyo.

Bivand, R. & Rundel, C. (2014). rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.3-2.

Bloomquist, E. W., Lemey, P., & Suchard, M. A. (2010). Three roads diverged? Routes to phylogeographic inference. *Trends in Ecology & Evolution*, **25**(11); pages 626–632. ISSN 0169-5347.

Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011). Apcluster: An R package for affinity propagation clustering. *Bioinformatics*, **27**(17); pages 2463–2464. ISSN 13674803.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15); pages 2114–2120. ISSN 14602059.

Bourhy, H., Kissi, B., Audry, L., Smreczak, M., Sadkowska-Todys, M., Kulonen, K., Tordo, N., Zmudzinski, J. F., & Holmes, E. C. (1999). Ecology and evolution of rabies virus in Europe. *The Journal of general virology*, **80 (Pt 10**(10); pages 2545–57. ISSN 0022-1317.

Bourhy, H., Reynes, J.-M., Dunham, E. J., Dacheux, L., Larrous, F., Huong, V. T. Q., Xu, G., Yan, J., Miranda, M. E. G., & Holmes, E. C. (2008). The origin and phylogeography of dog rabies virus. *The Journal of general virology*, **89**(Pt 11); pages 2673–81. ISSN 0022-1317.

Brien, J. D. O., Minin, V. N., & Suchard, M. A. (2009). Learning to Count : Robust Estimates for Labeled Distances between Molecular Sequences. *Molecular Biology and Evolution*, **26**(4); pages 801–814.

Brunker, K., Hampson, K., Horton, D. L., & Biek, R. (2012). Integrating the landscape epidemiology and genetics of RNA viruses: rabies in domestic dogs as a model. *Parasitology*, **139**(14); pages 1899–1913. ISSN 0031-1820.

Brunker, K., Marston, D. A., Horton, D. L., Cleaveland, S., Fooks, A. R., Kazwala, R., Ngeleja, C., Lembo, T., Sambo, M., Mtema, Z. J., Sikana, L., Wilkie, G., Biek, R., & Hampson, K. (2015). Elucidating the phylodynamics of endemic rabies virus in eastern Africa using whole-genome sequencing. *Virus Evolution*, **1**(1); page vev011. ISSN 2057-1577.

Burr, T., Doak, J., & Gattiker, J. (2002). Assessing confidence in phylogenetic trees: bootstrap versus Markov Chain Monte Carlo. *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences*, **836**(November).

Butler, J. R. & Bingham, J. (2000). Demography and dog-human relationships of the dog population in Zimbabwean communal lands. *The Veterinary record*, **147**(16); pages 442–6. ISSN 0042-4900.

Carnieli, P., de Novaes Oliveira, R., Macedo, C. I., & Castilho, J. G. (2011). Phylogeography of rabies virus isolated from dogs in Brazil between 1985 and 2006. *Archives of virology*, **156**(6); pages 1007–12. ISSN 1432-8798.

Carrington, C., Foster, J., Pybus, O. G., Bennett, S. N., & Holmes, E. C. (2005). Invasion and Maintenance of Dengue Virus Type 2 and Type 4 in the Americas. *Journal of ...*, **79**(23); pages 14,680–14,687.

Carvalho, L. M., Faria, N. R., Perez, A. M., Suchard, M. a., Lemey, P., Silveira, W. D. C., Rambaut, A., & Baele, G. (2015). Spatio-temporal Dynamics of Foot-and-Mouth Disease Virus in South America. *arXiv preprint arXiv:1505.01105*, pages 1–21.

Chare, E. R., Gould, E. A., & Holmes, E. C. (2003). Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. *Journal of General Virology*, **84**(10); pages 2691–2703. ISSN 00221317.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, **61**(6); pages 1–36.

Cleaveland, S. (1998). The growing problem of rabies in Africa.

Cleaveland, S., Haydon, D., & Taylor, L. (2007). Overviews of pathogen emergence: which pathogens emerge, when and why? *Current Topics in Microbiology and Immunology*, **315**; pages 85–111.

Cleaveland, S., Kaare, M., Tiringa, P., Mlengeya, T., & Barrat, J. (2003). A dog rabies vaccination campaign in rural Africa: impact on the incidence of dog rabies and human dog-bite injuries. *Vaccine*, **21**(17-18); pages 1965–1973.

Coetzee, P. & Nel, L. H. (2007). Emerging epidemic dog rabies in coastal South Africa: a molecular epidemiological analysis. *Virus research*, **126**(1-2); pages 186–95. ISSN 0168-1702.

Cohen, C., Sartorius, B., Sabeta, C., Zulu, G., Paweska, J., Mogoswane, M., Sutton, C., Nel, L. H., Swanepoel, R., Leman, P. A., Grobbelaar, A. a., Dyason, E., & Blumberg, L. (2007). Epidemiology and Molecular Virus Characterization of Reemerging Rabies, South Africa. *Emerging Infectious Diseases*, **13**(12); pages 1879–1886. ISSN 1080-6040.

Cottam, E. M., Wadsworth, J., Shaw, A. E., Rowlands, R. J., Goatley, L., Maan, S., Maan, N. S., Mertens, P. P. C., Ebert, K., Li, Y., Ryan, E. D., Juleff, N., Ferris, N. P., Wilesmith, J. W., Haydon, D. T., King, D. P., Paton, D. J., & Knowles, N. J. (2008). Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007. *PLoS pathogens*, **4**(4); page e1000,050. ISSN 1553-7374.

Cross, P. C., Johnson, P. L. F., Lloyd-Smith, J. O., & Getz, W. M. (2007). Utility of R0 as a predictor of disease invasion in structured populations. *Journal of the Royal Society, Interface / the Royal Society*, **4**(13); pages 315–24. ISSN 1742-5689.

Cross, P. C., Lloyd-Smith, J. O., Johnson, P. L. F., & Getz, W. M. (2005). Duelling timescales of host movement and disease recovery determine invasion of disease in structured populations. *Ecology Letters*, **8**(6); pages 587–595. ISSN 1461023X.

Cullingham, C. I., Kyle, C. J., Pond, B. A., Rees, E. E., & White, B. N. (2009). Differential permeability of rivers to raccoon gene flow corresponds to rabies incidence in Ontario, Canada. *Molecular ecology*, **18**(1); pages 43–53. ISSN 1365-294X.

Cullingham, C. I., Kyle, C. J., Pond, B. A., & White, B. N. (2008). Genetic structure of raccoons in eastern North America based on mtDNA: implications for subspecies designation and rabies disease dynamics. *Canadian Journal of Zoology*, **86**(9); pages 947–958. ISSN 0008-4301.

David, D., Hughes, G. J., Yakobson, B. a., Davidson, I., Un, H., Aylan, O., Kuzmin, I. V., & Rupprecht, C. E. (2007). Identification of novel canine rabies virus clades in the Middle East and North Africa. *The Journal of general virology*, **88**(Pt 3); pages 967–80. ISSN 0022-1317.

David, D., Yakobson, B. A., Gershkovich, L., & Gayer, S. (2004). Tracing the regional source of rabies infection in an Israeli dog by viral analysis. *Veterinary Record*, **155**(16); pages 496–497. ISSN 0042-4900.

De Mattos, C. C., De Mattos, C. A., Loza-Rubio, E., Aguilar-Setién, A., Orciari, L. A., & Smith, J. S. (1999). Molecular characterization of rabies virus isolates from Mexico: implications for transmission dynamics and human risk. *The American journal of tropical medicine and hygiene*, **61**(4); pages 587–97. ISSN 0002-9637.

Dellicour, S., Michez, D., Rasplus, J.-Y., & Mardulyn, P. (2015). Impact of past climatic changes and resource availability on the population demography of three food-specialist bees. *Molecular ecology*. ISSN 1365-294X.

Dellicour, S., Rose, R., & Pybus, O. G. (2016). Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics*, **17**(1); pages 1–12. ISSN 1471-2105.

Denduangboripant, J., Wacharapluesadee, S., Lumlertdacha, B., Ruankaew, N., Hoonsuwan, W., Puanghat, A., & Hemachudha, T. (2005). Transmission dynamics of rabies virus in Thailand: implications for disease control. *BMC infectious diseases*, **5**; page 52. ISSN 1471-2334.

Dietzschold, B., Schnell, M., & Koprowski, H. (2005). Pathogenesis of rabies. *Current topics in microbiology and immunology*, **292**; pages 45–56. ISSN 0070-217X.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, **4**(5); page e88.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**(3); pages 1307–1320. ISSN 00166731.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., & Rodrigo, A. G. (2003). Measurably evolving populations. *Trends in Ecology & Evolution*, **18**(9); pages 481–488. ISSN 01695347.

Drummond, A. J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, **7**; page 214. ISSN 1471-2148.

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, **22**(5); pages 1185–92. ISSN 0737-4038.

Drummond, A. J., Suchard, M. a., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*, **29**(8); pages 1969–73. ISSN 1537-1719.

Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature reviews. Genetics*, **9**(4); pages 267–76. ISSN 1471-0064.

Eng, T. R., Fishbein, D. B., Talamante, H. E., Hall, D. B., Chavez, G. F., Dobbins, J. G., Muro, F. J., Bustos, J. L., de los Angeles Ricardy, M., & Munguia, A. (1993). Urban epizootic of rabies in Mexico: epidemiology and impact of animal bite injuries. *Bulletin of the World Health Organization*, **71**(5); pages 615–24. ISSN 0042-9686.

Epperson, B. K., McRae, B. H., Scribner, K., Cushman, S. A., Rosenberg, M. S., Fortin, M. J., James, P. M. A., Murphy, M., Manel, S., Legendre, P., & Dale, M. R. T. (2010). Utility of computer simulations in landscape genetics. *Molecular Ecology*, **19**(17); pages 3549–3564. ISSN 09621083.

Etherington, T. R. (2011). Python based GIS tools for landscape genetics: visualising genetic relatedness and measuring landscape connectivity. *Methods in Ecology and Evolution*, **2**(1); pages 52–55. ISSN 2041210X.

Excoffier, L. & Ray, N. (2008). Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**(7); pages 347–351. ISSN 01695347.

Faria, N. R., Suchard, M. A., Rambaut, A., & Lemey, P. (2011). Toward a quantitative understanding of viral phylogeography. *Current Opinion in Virology*, **1**.

Ferreira, M. a. R. & Suchard, M. a. (2008). Bayesian analysis of elapsed times in continuous-time Markov chains. *Canadian Journal of Statistics*, **36**(3); pages 355–368. ISSN 03195724.

Fraley, C. & Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation.

Frey, B. J. & Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, **315**(5814); pages 972–976. ISSN 0036-8075.

Frost, S. D. W., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S., & Bedford, T. (2015). Eight challenges in phylodynamic inference. *Epidemics*, **10**; pages 88–92. ISSN 1878-0067.

Frost, S. D. W. & Volz, E. M. (2013). Modelling tree shape and structure in viral phylodynamics. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **368**(1614); page 20120,208. ISSN 1471-2970.

Gardy, J. L., Johnston, J. C., Ho Sui, S. J., Cook, V. J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., Varhol, R., Birol, I., Lem, M., Sharma, M. K., Elwood, K., Jones, S. J. M., Brinkman, F. S. L., Brunham, R. C., & Tang, P. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *The New England journal of medicine*, **364**(8); pages 730–9. ISSN 1533-4406.

Gautret, P., Ribadeau-Dumas, F., Parola, P., Brouqui, P., & Bourhy, H. (2011). Risk for rabies importation from North Africa. *Emerging Infectious Diseases*, **17**(12); pages 2187–2193. ISSN 10806040.

Greger, M. (2007). The human/animal interface: emergence and resurgence of zoonotic infectious diseases. *Critical reviews in microbiology*, **33**(4); pages 243–99. ISSN 1040-841X.

Grenfell, B. & Harwood, J. (1997). (Meta) population dynamics of infectious diseases. *Trends in Ecology & Evolution*, **12**(10); pages 395–399.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. a., & Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, N.Y.)*, **303**(5656); pages 327–32. ISSN 1095-9203.

Hagenaars, T. J., Donnelly, C. a., & Ferguson, N. M. (2004). Spatial heterogeneity and the persistence of infectious diseases. *Journal of theoretical biology*, **229**(3); pages 349–59. ISSN 0022-5193.

Hall, M., Woolhouse, M., & Rambaut, A. (2015). Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLOS Computational Biology*, **11**(12); page e1004,613. ISSN 1553-7358.

Hall, M. D., Woolhouse, M. E. J., & Rambaut, A. (2016). The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods : A simulation study. *Virus Evolution*, **2**(1); pages 1–14.

Halliday, J. E. B., Allan, K. J., Ekwem, D., Cleaveland, S., Kazwala, R. R., & Crump, J. A. (2015). Endemic zoonoses in the tropics: a public health problem hiding in plain sight. *Veterinary Record*, **176**(9); pages 220–225. ISSN 0042-4900.

Hampson, K., Dushoff, J., Bingham, J., Brückner, G., Ali, Y. H., & Dobson, A. (2007). Synchronous cycles of domestic dog rabies in sub-Saharan Africa and the impact of control efforts. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(18); pages 7717–22. ISSN 0027-8424.

Hampson, K., Dushoff, J., Cleaveland, S., Haydon, D. T., Kaare, M., Packer, C., & Dobson, A. (2009). Transmission Dynamics and Prospects for the Elimination of Canine Rabies. *PLoS Biology*, **7**(3); page e53. ISSN 1545-7885.

Han, G. Z. & Worobey, M. (2011). Homologous recombination in negative sense RNA viruses. *Viruses*, **3**(8); pages 1358–1373. ISSN 19994915.

Hanlon, C., Niezgoda, M., & Rupprecht, C. (2007). Rabies in terrestrial animals. *Rabies*.

Haydon, D. T., Chase-Topping, M., Shaw, D. J., Matthews, L., Friar, J. K., Wilesmith, J., & Woolhouse, M. E. J. (2003). The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proceedings. Biological sciences / The Royal Society*, **270**(1511); pages 121–7. ISSN 0962-8452.

Haydon, D. T., Randall, D. A., Matthews, L., Knobel, D. L., Tallents, L. A., Gravenor, M. B., Williams, S. D., Pollinger, J. P., Cleaveland, S., Woolhouse, M. E. J., Sillero-Zubiri, C., Marino, J., Macdonald, D. W., & Laurenson, M. K. (2006). Low-coverage vaccination strategies for the conservation of endangered species. *Nature*, **443**(7112); pages 692–695. ISSN 0028-0836.

Hayman, D. T. S., Johnson, N., Horton, D. L., Hedge, J., Wakeley, P. R., Banyard, A. C., Zhang, S., Alhassan, A., & Fooks, A. R. (2011). Evolutionary history of rabies in Ghana. *PLoS neglected tropical diseases*, **5**(4); page e1001. ISSN 1935-2735.

Heaton, P. R., Johnstone, P., McElhinney, L. M., Cowley, R., O'Sullivan, E., & Whitby, J. E. (1997). Heminested PCR assay for detection of six genotypes of rabies and rabies-related viruses. *Journal of Clinical Microbiology*. ISSN 00951137.

Hennig, C. (2014). fpc: Flexible procedures for clustering.

Hillis, D. M. & Bull, J. J. (1993). An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology*, **42**(2); pages 182–192. ISSN 1063-5157.

Ho, S. Y. W. & Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular ecology resources*, **11**(3); pages 423–34. ISSN 1755-0998.

Holder, M. & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews. Genetics*, **4**(4); pages 275–284. ISSN 14710056.

Holmes, E. C. (2004). The phylogeography of human viruses.

Holmes, E. C. & Grenfell, B. T. (2009). Discovering the phylodynamics of RNA viruses. *PLoS computational biology*, **5**(10); page e1000,505. ISSN 1553-7358.

Holmes, E. C. & Moya, A. (2002). Is the quasispecies qoncept relevant to RNA viruses? *Journal of Virology*, **76**(1); pages 460–462.

Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P., & Harvey, P. H. (1995). Revealing the history of infectious disease epidemics through phylogenetic trees. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **349**(1327); pages 33–40. ISSN 0962-8436.

Holmes, E. C., Woelk, C. H., Kassis, R., & Bourhy, H. (2002). Genetic constraints and the adaptive evolution of rabies virus in nature. *Virology*, **292**(2); pages 247–257. ISSN 0042-6822.

Horton, D. L., McElhinney, L. M., Freuling, C. M., Marston, D. a., Banyard, A. C., Goharrriz, H., Wise, E., Breed, A. C., Saturday, G., Kolodziejek, J., Zilahi, E., Al-Kobaisi, M. F., Nowotny, N., Mueller, T., & Fooks, A. R. (2015). Complex Epidemiology of a Zoonotic Disease in a Culturally Diverse Region: Phylogeography of Rabies Virus in the Middle East. *PLOS Neglected Tropical Diseases*, **9**; page e0003,569. ISSN 1935-2735.

Huelsenbeck, J. & Rannala, B. (2004). Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic biology*, **53**(6); pages 904–913. ISSN 1063-5157.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science (New York, N.Y.)*, **294**(5550); pages 2310–2314. ISSN 00368075.

Hughes, J., Allen, R. C., Baguelin, M., Hampson, K., Baillie, G. J., Elton, D., Newton, J. R., Kellam, P., Wood, J. L. N., Holmes, E. C., & Murcia, P. R. (2012). Transmission of equine influenza virus during an outbreak is characterized by frequent mixed infections and loose transmission bottlenecks. *PLoS pathogens*, **8**(12); page e1003,081. ISSN 1553-7374.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics (Oxford, England)*, **24**(11); pages 1403–5. ISSN 1367-4811.

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., & Ferguson, N. (2014). Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology*, **10**(1). ISSN 1553734X.

Jones, K. E., Patel, N. G., Levy, M. a., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, **451**(7181); pages 990–3. ISSN 1476-4687.

Kao, R. R., Haydon, D. T., Lycett, S. J., & Murcia, P. R. (2014). Supersize me: how whole-genome sequencing and big data are transforming epidemiology. *Trends in microbiology*, **22**(5); pages 282–291. ISSN 1878-4380.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430); pages 773–795.

Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**(4); pages 772–80. ISSN 1537-1719.

Keeling, M. J., Woolhouse, M. E. J., May, R. M., Davies, G., & Grenfell, B. T. (2003). Modelling vaccination strategies against foot-and-mouth disease. *Nature*, **421**(6919); pages 136–142. ISSN 0028-0836.

Kideghesho, J., Rija, A., Mwamende, K., & Selemani, I. (2013). Emerging issues and challenges in conservation of biodiversity in the rangelands of Tanzania. *Nature Conservation*, **6**; pages 1–29. ISSN 1314-3301.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **78**(1); pages 454–458. ISSN 0027-8424.

King, A. A., of Epizootics, I. O., for the Characterisation of Rabies, W. H. O. C. C., Viruses, R.-r., & Agency, V. L. (2004). *Historical perspective of rabies in Europe and the Mediterranean Basin: a testament to rabies by Dr. Arthur A. King.* World Organisation for Animal Health. ISBN 9789290446392.

Kiskowski, M. & Chowell, G. (2015). Modeling household and community transmission of Ebola virus disease: epidemic growth, spatial dynamics and insights for epidemic control. *Virulence*, **(just-acce**(September 2015); pages 1–11. ISSN 2150-5594.

Kissi, B., Tordo, N., & Bourhy, H. (1995). Genetic Polymorphism in the Rabies Virus Nucleoprotein Gene. *Virology*, **209**(2); pages 526–537. ISSN 00426822.

Kitala, P., McDermott, J., Kyule, M., Gathuma, J., Perry, B., & Wandeler, A. (2001). Dog ecology and demography information to support the planning of rabies control in Machakos District, Kenya. *Acta Tropica*, **78**(3); pages 217–230. ISSN 0001706X.

Klepac, P., Metcalf, C. J. E., McLean, A. R., & Hampson, K. (2013). Towards the endgame and beyond: complexities and challenges for the elimination of infectious diseases. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **368**(1623); page 20120,137. ISSN 1471-2970.

Klopfstein, S., Currat, M., & Excoffier, L. (2006). The fate of mutations surfing on the wave of a range expansion. *Molecular biology and evolution*, **23**(3); pages 482–90. ISSN 0737-4038.

Knobel, D. L., Cleaveland, S., Coleman, P. G., Fèvre, E. M., Meltzer, M. I., Miranda, M. E. G., Shaw, A., Zinsstag, J., & Meslin, F.-X. (2005). Re-evaluating the burden of rabies in Africa and Asia. *Bulletin of the World Health Organization*, **83**(5); pages 360–8. ISSN 0042-9686.

Knobel, D. L., Laurenson, M. K., Kazwala, R. R., Boden, L. A., & Cleaveland, S. (2008). A cross-sectional study of factors associated with dog ownership in Tanzania. *BMC veterinary research*, **4**; page 5. ISSN 1746-6148.

Kühnert, D., Wu, C.-H., & Drummond, A. J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, **11**(8); pages 1825–41. ISSN 1567-7257.

Kuzmin, I. V., Wu, X., Tordo, N., & Rupprecht, C. E. (2008). Complete genomes of Aravan, Khujand, Irkut and West Caucasian bat viruses, with special attention to the polymerase gene and non-coding regions. *Virus research*, **136**(1-2); pages 81–90. ISSN 0168-1702.

Kuzmina, N. a., Lemey, P., Kuzmin, I. V., Mayes, B. C., Ellison, J. a., Orciari, L. a., Hightower, D., Taylor, S. T., & Rupprecht, C. E. (2013). The Phylogeography and Spatiotemporal Spread of South-Central Skunk Rabies Virus. *PLoS ONE*, **8**(12); page e82,348. ISSN 1932-6203.

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular biology and evolution*, **29**(6); pages 1695–701. ISSN 1537-1719.

Lembo, T., Attlan, M., Bourhy, H., Cleaveland, S., Costa, P., de Balogh, K., Dodet, B., Fooks, A. R., Hiby, E., Leanes, F., Meslin, F.-X., Miranda, M. E., Müller, T., Nel, L. H., Rupprecht, C. E., Tordo, N., Tumpey, A., Wandeler, A., & Briggs, D. J. (2011). Renewed global partnerships and redesigned roadmaps for rabies prevention and control. *Veterinary medicine international*, **2011**; page 923,149. ISSN 2042-0048.

Lembo, T., Hampson, K., Haydon, D. T., Craft, M., Dobson, A., Dushoff, J., Ernest, E., Hoare, R., Kaare, M., Mlengeya, T., Mentzel, C., & Cleaveland, S. (2008). Exploring reservoir dynamics: a case study of rabies in the Serengeti ecosystem. *Journal of Applied Ecology*, **45**(4); pages 1246–1257. ISSN 00218901.

Lembo, T., Hampson, K., Kaare, M. T., Ernest, E., Knobel, D., Kazwala, R. R., Haydon, D. T., Cleaveland, S., & Rudovick, R. (2010). The feasibility of canine rabies elimination in Africa: dispelling doubts with data. *PLoS neglected tropical diseases*, **4**(2); page e626. ISSN 1935-2735.

Lembo, T., Haydon, D. T., Velasco-Villa, A., Rupprecht, C. E., Packer, C., Brandão, P. E., Kuzmin, I. V., Fooks, A. R., Barrat, J., & Cleaveland, S. (2007). Molecular epidemiology identifies only a single rabies virus variant circulating in complex carnivore communities of the Serengeti. *Proceedings. Biological sciences / The Royal Society*, **274**(1622); pages 2123–30. ISSN 0962-8452.

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. a., Smith, D. J., Pybus, O. G., Brockmann, D., & Suchard, M. a. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS pathogens*, **10**(2); page e1003,932. ISSN 1553-7374.

Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. a. (2009). Bayesian phylogeography finds its roots. *PLoS computational biology*, **5**(9); page e1000,520. ISSN 1553-7358.

Lemey, P., Rambaut, A., & Pybus, O. G. (2006). HIV evolutionary dynamics within and among hosts. *AIDS Reviews*, **8**(3); pages 125–140. ISSN 11396121.

Lemey, P., Rambaut, A., Welch, J. J., & Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol*, **27**.

Levin, S. A. (1992). The problem of pattern and scale in ecology.

Lewin-Koh, N. J., Bivand, R., Pebesma, E. J., Archer, E., Baddeley, A., Dray, S., Forrest, D., Giraudoux, P., Golicher, D., Rubio, G. Ã., Hausmann, P., Jagger, T., Luque, S. P., Macqueen, D., Niccolai, A., Short, T., & Rogerbivandnhhno, M. R. B. (2012). Maptools: Tools for reading and handling spatial objects. *R package version 0.8-14*.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **25**(14); pages 1754–60. ISSN 1367-4811.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**(16); pages 2078–9. ISSN 1367-4811.

Liu, W., Liu, Y., Liu, J., Zhai, J., & Xie, Y. (2011). Evidence for inter- and intra-clade recombinations in rabies virus. *Infection, Genetics and Evolution*, **11**(8); pages 1906–1912. ISSN 15671348.

Lloyd, a. L. & May, R. M. (1996). Spatial heterogeneity in epidemic models. *Journal of theoretical biology*, **179**(1); pages 1–11. ISSN 0022-5193.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, **438**(7066); pages 355–9. ISSN 1476-4687.

Lu, L., Lycett, S. J., & Leigh Brown, A. J. (2014). Determining the phylogenetic and phylogeographic origin of highly pathogenic avian influenza (H7N3) in Mexico. *PloS one*, **9**(9); page e107,330. ISSN 1932-6203.

Lumlertdacha, B., Wacharapluesadee, S., Denduangboripant, J., Ruankaew, N., Hoonsuwan, W., Puanghat, A., Sakarasaeranee, P., Briggs, D., & Hemachudha, T. (2006). Complex genetic structure of the rabies virus in Bangkok and its surrounding provinces, Thailand: implications for canine rabies control. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **100**(3); pages 276–281.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2015). Cluster Analysis Basics and Extensions. R package version 2.0.1.

Magee, D., Beard, R., Suchard, M. a., Lemey, P., & Scotch, M. (2014). Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza A virus diffusion. *Archives of Virology*, **160**; pages 215–224. ISSN 0304-8608.

Magembe, S. (1985). Epidemiology of Rabies in United Republic of Tanzania. In Kuwer, E., MeÌĄrieux, C., Koprowski, H., & BoÌĹgel, K., editors, *Rabies in the tropics*. Springer Berlin Heidelberg New York.

Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**(4); pages 189–197. ISSN 01695347.

Marston, D. a., McElhinney, L. M., Ellis, R. J., Horton, D. L., Wise, E. L., Leech, S. L., David, D., de Lamballerie, X., & Fooks, A. R. (2013). Next generation sequencing of viral RNA genomes. *BMC genomics*, **14**(1); page 444. ISSN 1471-2164.

Marston, D. a., McElhinney, L. M., Johnson, N., Müller, T., Conzelmann, K. K., Tordo, N., & Fooks, a. R. (2007). Comparative analysis of the full genome sequence of European bat lyssavirus type 1 and type 2 with other lyssaviruses and evidence for a conserved transcription termination and polyadenylation motif in the G-L 3' non-translated region. *The Journal of general virology*, **88**(Pt 4); pages 1302–14. ISSN 0022-1317.

Matsumoto, T., Ahmed, K., Karunanayake, D., Wimalaratne, O., Nanayakkara, S., Perera, D., Kobayashi, Y., & Nishizono, A. (2013). Molecular epidemiology of human rabies viruses in Sri Lanka. *Infection, Genetics and Evolution*, **18**; pages 160–167. ISSN 15671348.

Maudlin, I., Eisler, M. C., & Welburn, S. C. (2009). Neglected and endemic zoonoses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **364**(1530); pages 2777–2787. ISSN 0962-8436.

McCallum, H. (2008). Landscape Structure, Disturbance, and Disease Dynamics. In Richard S Ostfeld;, Felicia Keesing;, & Valerie, T. E., editors, *Infectious disease ecology : the effects of ecosystems on disease and of disease on ecosystems*, chapter Five, pages 100–122. Princeton University Press, Princeton, New Jersey.

McElhinney, L. M., Marston, D. a., Freuling, C. M., Cragg, W., Stankov, S., Lalosevic, D., Lalosevic, V., Müller, T., & Fooks, a. R. (2011). Molecular diversity and evolutionary history of rabies virus strains circulating in the Balkans. *The Journal of general virology*, **92**(Pt 9); pages 2171–80. ISSN 1465-2099.

McRae, B., Dickson, B., Keitt, T., & Shah, V. (2008). Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, **89**(10); pages 2712–2724.

McRae, B. H. (2006). Isolation by resistance. *Evolution*, **60**(8); pages 1551–1561. ISSN 0014-3820.

McRae, B. H. & Beier, P. (2007). Circuit theory predicts gene flow in plant and animal populations. *Proc Natl Acad Sci U S A*, **104**.

Meentemeyer, R. K., Cunniffe, N. J., Cook, A. R., Filipe, J. a. N., Hunter, R. D., Rizzo, D. M., & Gilligan, C. a. (2011). Epidemiological modeling of invasion in heterogeneous landscapes: spread of sudden oak death in California (1990âĂŞ2030). *Ecosphere*, **2**(2); page art17. ISSN 2150-8925.

Meentemeyer, R. K., Haas, S. E., & Václavík, T. (2012). Landscape Epidemiology of Emerging Infectious Diseases in Natural and Human-Altered Ecosystems. *Annual review of phytopathology*, (May). ISSN 0066-4286.

Metcalf, C., Hampson, K., & Tatem, A. (2013). Persistence in epidemic metapopulations: quantifying the rescue effects for measles, mumps, rubella and whooping cough. *PloS one*, **8**(9); page e74,696. ISSN 1932-6203.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, **11**(1); pages 31–46. ISSN 1471-0064.

Minin, V. N. & Suchard, M. a. (2008). Counting labeled transitions in continuous-time Markov models of evolution. *Journal of mathematical biology*, **56**(3); pages 391–412. ISSN 1432-1416.

Mollentze, N., Biek, R., & Streicker, D. G. (2014a). The role of viral evolution in rabies host shifts and emergence. *Current Opinion in Virology*, **8**; pages 68–72. ISSN 18796257.

Mollentze, N., Nel, L. H., Townsend, S., le Roux, K., Hampson, K., Haydon, D. T., & Soubeyrand, S. (2014b). A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings of the Royal Society B: Biological Sciences*, **281**(1782); pages 20133,251–20133,251. ISSN 0962-8452.

Mollentze, N., Weyer, J., Markotter, W., le Roux, K., & Nel, L. H. (2013). Dog rabies in southern Africa: regional surveillance and phylogeographical analyses are an important component of control and elimination strategies. *Virus genes*, **47**(3); pages 569–73. ISSN 1572-994X.

Morelli, M. J., Thébaud, G., Chadœuf, J., King, D. P., Haydon, D. T., & Soubeyrand, S. (2012). A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS computational biology*, **8**(11); page e1002,768. ISSN 1553-7358.

Morelli, M. J., Wright, C. F., Knowles, N. J., Juleff, N., Paton, D. J., King, D. P., & Haydon, D. T. (2013). Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. *Veterinary research*, **44**(1); page 12. ISSN 0928-4249.

Morters, M. K., Restif, O., Hampson, K., Cleaveland, S., Wood, J. L., & Conlan, A. J. (2013). Evidence-based control of canine rabies: a critical review of population density reduction. *Journal of Animal Ecology*, **82**(1); pages 6–14. ISSN 1365-2656.

Nadin-davis, S. A., Feng, Y., Mousse, D., & Wandeler, A. I. (2010). Spatial and temporal dynamics of rabies virus variants in big brown bat populations across Canada : footprints of an emerging zoonosis. *Molecular ecology*, **19**; pages 2120–2136.

Nadin-Davis, S. a., Sampath, M. I., Casey, G. a., Tinline, R. R., & Wandeler, a. I. (1999). Phylogeographic patterns exhibited by Ontario rabies virus variants. *Epidemiology and infection*, **123**(2); pages 325–36. ISSN 0950-2688.

Nel, L. H. (2013). Discrepancies in data reporting for rabies, Africa. *Emerging infectious diseases*, **19**(4); pages 529–33. ISSN 1080-6059.

Nelson, M. I., Viboud, C., Vincent, A. L., Culhane, M. R., Detmer, S. E., Wentworth, D. E., Rambaut, A., Suchard, M. a., Holmes, E. C., & Lemey, P. (2015). Global migration of influenza A viruses in swine. *Nature Communications*, **6**; page 6696. ISSN 2041-1723.

Nettles, V. F., Shaddock, J. H., Sikes, R. K., & Reyes, C. R. (1979). Rabies in translocated raccoons. *American Journal of Public Health*, **69**(6); pages 601–602.

Nunes, M. R. T., Palacios, G., Faria, N. R., Sousa, E. C., Pantoja, J. a., Rodrigues, S. G., Carvalho, V. L., Medeiros, D. B. a., Savji, N., Baele, G., Suchard, M. a., Lemey, P., Vasconcelos, P. F. C., & Lipkin, W. I. (2014). Air travel is associated with intracontinental

spread of dengue virus serotypes 1-3 in Brazil. *PLoS neglected tropical diseases*, **8**(4); page e2769. ISSN 1935-2735.

Oh, M.-S. & Raftery, A. E. (2001). Bayesian Multidimensional Scaling and Choice of Dimension. *Journal of the American Statistical Association*, **96**(455); pages 1031–1044. ISSN 0162-1459.

Orton, R. J., Wright, C. F., Morelli, M. J., Juleff, N., Thébaud, G., Nick, J., Valdazo-gonzález, B., Paton, D. J., King, D. P., Haydon, D. T., B, P. T. R. S., & Knowles, N. J. (2013). Observing micro-evolutionary processes of viral populations at multiple scales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**(1614).

Ospina, M. C., Diaz, F. J., & Osorio, J. E. (2010). Prolonged co-circulation of two distinct dengue virus type 3 lineages in the hyperendemic area of Medellín, Colombia. *American Journal of Tropical Medicine and Hygiene*, **83**(3); pages 672–678. ISSN 00029637.

Ostfeld, R. S., Glass, G. E., & Keesing, F. (2005). Spatial epidemiology: an emerging (or re-emerging) discipline. *Trends in ecology & evolution*, **20**(6); pages 328–36. ISSN 0169-5347.

Paetkau, D., Calvert, W., Stirling, I., & Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular ecology*, **4**(3); pages 347–354. ISSN 0962-1083.

Pagel, M., Meade, A., & Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic biology*, **53**(5); pages 673–684. ISSN 1063-5157.

Panjeti, V. G. & Real, L. a. (2011). *Mathematical models for rabies.*, volume 79. Elsevier Inc., 1 edition. ISBN 9780123870407.

Paradis E., C. J. &. S. K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**; pages 289–290.

Parker, J., Rambaut, A., & Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution*, **8**(3); pages 239–246. ISSN 15671348.

Pavlovsky, E. N. & Levine, N. D. (1966). *Natural Nidality of Transmissible Diseases: With Special Reference to the Landscape Epidemiology of Zooanthroponoses*. University of IllinoisPress. ISBN 9780608137414.

Pebesma, E. & Bivand, R. (2005). Classes and methods for spatial data in R. Technical report.

Pfeiffer, D. U. & Stevens, K. B. (2015). Spatial and temporal epidemiological analysis in the Big Data era. *Preventive Veterinary Medicine*, **0**. ISSN 01675877.

PINGO's Forum (2013). Pastoralists Indigenous Non Governmental Organizations' Forum Annual Report 2012-2013. Technical Report November 2012.

Pulliam, H. R. (1988). Sources Sinks and Population Regulation. *American Naturalist*, **132**(5); pages 652–661. ISSN 0003-0147.

Pybus, O. G. & Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature reviews. Genetics*, **10**(8); pages 540–550. ISSN 1471-0056.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., & Delwart, E. L. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics.

R Core Team (2015). R: A language and environment for statistical computing.

Raghwani, J., Rambaut, A., Holmes, E. C., Hang, V. T., Hien, T. T., Farrar, J., Wills, B., Lennon, N. J., Birren, B. W., Henn, M. R., & Simmons, C. P. (2011). Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission. *PLoS Pathogens*, **7**(6); pages 1–10. ISSN 15537366.

Rambaut, A. & Drummond, A. J. (2014). Tracer V1.6.

Rasmussen, D. a., Ratmann, O., & Koelle, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Computational Biology*, **7**(8). ISSN 1553734X.

Real, L. A. & Biek, R. (2007). Spatial dynamics and genetics of infectious diseases on heterogeneous landscapes. *J R Soc Interface*, **4**.

Real, L. a., Henderson, J. C., Biek, R., Snaman, J., Jack, T. L., Childs, J. E., Stahl, E., Waller, L., Tinline, R., & Nadin-Davis, S. (2005a). Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(34); pages 12,107–11. ISSN 0027-8424.

Real, L. a., Russell, C., Waller, L., Smith, D., & Childs, J. (2005b). Spatial dynamics and molecular ecology of North American rabies. *The Journal of heredity*, **96**(3); pages 253–60. ISSN 0022-1503.

Rees, E. E., Pond, B. A., Cullingham, C. I., Tinline, R., Ball, D., Kyle, C. J., & White, B. N. (2008). Assessing a landscape barrier using genetic simulation modelling: implications for raccoon rabies management. *Preventive veterinary medicine*, **86**(1-2); pages 107–23. ISSN 0167-5877.

Rees, E. E., Pond, B. A., Cullingham, C. I., Tinline, R. R., Ball, D., Kyle, C. J., & White, B. N. (2009). Landscape modelling spatial bottlenecks: implications for raccoon rabies disease spread. *Biology letters*, **5**(3); pages 387–90. ISSN 1744-9561.

Rees, E. E., Pond, B. a., Tinline, R. R., & Bélanger, D. (2013). Modelling the effect of landscape heterogeneity on the efficacy of vaccination for wildlife infectious disease control. *Journal of Applied Ecology*, **50**(4); pages 881–891. ISSN 00218901.

Rosenberg, M. S. & Anderson, C. D. (2011). PASSaGE: Pattern Analysis, Spatial Statistics and Geographic Exegesis. Version 2. *Methods in Ecology and Evolution*, **2**(3); pages 229–232. ISSN 2041210X.

Rupprecht, C. E., Hanlon, C. a., & Hemachudha, T. (2002). Rabies re-examined. *The Lancet infectious diseases*, **2**(6); pages 327–43. ISSN 1473-3099.

Russell, C. a., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M.,

Odagiri, T., Osterhaus, A. D. M. E., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. a. M., & Smith, D. J. (2008). The global circulation of seasonal influenza A (H3N2) viruses. *Science (New York, N.Y.)*, **320**(5874); pages 340–6. ISSN 1095-9203.

Russell, C. A., Real, L. A., & Smith, D. L. (2006). Spatial Control of Rabies on Heterogeneous Landscapes. *PLoS ONE*, **1**(1); page 7. ISSN 1932-6203.

Russell, C. a., Smith, D. L., Childs, J. E., & Real, L. a. (2005). Predictive spatial dynamics and strategic planning for raccoon rabies emergence in Ohio. *PLoS biology*, **3**(3); page e88. ISSN 1545-7885.

Russell, C. a., Smith, D. L., Waller, L. a., Childs, J. E., & Real, L. a. (2004). A priori prediction of disease invasion dynamics in a novel environment. *Proceedings. Biological sciences / The Royal Society*, **271**(1534); pages 21–5. ISSN 0962-8452.

Schneider, M. C. & Leanes, L. F. (2007). Current status of human rabies transmitted by dogs in Latin America Raiva humana transmitida por caninos : situação atual na América Latina. *Pan American Health*, **23**(9); pages 2049–2063.

Shah, V. & McRae, B. (2008). Circuitscape : A Tool for Landscape Ecology. In *Proceedings of the 7th Python in Science Conference*, volume 7, pages 62–65.

Shwiff, S., Hampson, K., & Anderson, A. (2013). Potential economic benefits of eliminating canine rabies. *Antiviral research*, **98**(2); pages 352–6. ISSN 1872-9096.

Siongok, T. & Karama, M. (1985). Epidemiology of human rabies in Kenya. In Kuwer, E., MeÌĄrieux, C., Koprowski, H., & BoÌĹgel, K., editors, *Rabies in the tropics*. Springer Berlin Heidelberg New York.

Smith, D. L., Lucey, B., Waller, L. A., Childs, J. E., & Real, L. A. (2002). Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. *Proc Natl Acad Sci U S A*, **99**.

Smith, D. L., Waller, L. a., Russell, C. a., Childs, J. E., & Real, L. a. (2005). Assessing the role of long-distance translocation and spatial heterogeneity in the raccoon rabies epidemic in Connecticut. *Preventive veterinary medicine*, **71**(3-4); pages 225–40. ISSN 0167-5877.

Smith, J. S., Orciari, L. A., Yager, P. A., Seidel, H. D., & Warner, C. K. (1992). Epidemiologic and historical relationships among 87 rabies virus isolates as determined by limited sequence analysis. *The Journal of infectious diseases*, **166**(2); pages 296–307. ISSN 0022-1899.

Soubeyrand, S. (2014). Construction of semi-Markov genetic-space-time SEIR models and inference. *<hal-01090675>*.

Spear, S. F., Balkenhol, N., Fortin, M. J., McRae, B. H., & Scribner, K. (2010). Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis. *Molecular Ecology*, **19**; pages 3576–3591. ISSN 09621083.

Stadler, T. (2009). On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, **261**(1); pages 58–66. ISSN 00225193.

Stadler, T. & Bonhoeffer, S. (2013). Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **368**(1614); page 20120,198. ISSN 1471-2970.

Stamatakis, a., Aberer, a. J., Goll, C., Smith, S. a., Berger, S. a., & Izquierdo-Carrasco, F. (2012). RAxML-Light: a tool for computing terabyte phylogenies. *Bioinformatics (Oxford, England)*, **28**(15); pages 2064–6. ISSN 1367-4811.

Susilawathi, N. M., Darwinata, A. E., Dwija, I. B., Budayanti, N. S., Wirasandhi, G. A., Subrata, K., Susilarini, N. K., Sudewi, R. A., Wignall, F. S., & Mahardika, G. N. (2012). Epidemiological and clinical features of human rabies cases in Bali 2008-2010. *BMC Infectious Diseases*, **12**(1); page 81. ISSN 1471-2334.

Swanepoel, R., Barnard, B. J., Meredith, C. D., Bishop, G. C., Brückner, G. K., Foggin, C. M., & Hübschle, O. J. (1993). Rabies in southern Africa. *The Onderstepoort journal of veterinary research*, **60**(4); pages 325–346. ISSN 0030-2465.

Szanto, A. G., Nadin-Davis, S. a., Rosatte, R. C., & White, B. N. (2011). Genetic tracking of the raccoon variant of rabies virus in eastern North America. *Epidemics*, **3**(2); pages 76–87. ISSN 1878-0067.

Talbi, C., Holmes, E. C., de Benedictis, P., Faye, O., Nakouné, E., Gamatié, D., Diarra, A., Elmamy, B. O., Sow, A., Adjogoua, E. V., Sangare, O., Dundon, W. G., Capua, I., Sall, A. a., & Bourhy, H. (2009). Evolutionary history and dynamics of dog rabies virus in western and central Africa. *The Journal of general virology*, **90**(Pt 4); pages 783–91. ISSN 0022-1317.

Talbi, C., Lemey, P., Suchard, M. a., Abdelatif, E., Elharrak, M., Nourlil, J., Jalal, N., Faouzi, A., Echevarría, J. E., Vazquez Morón, S., Rambaut, A., Campiz, N., Tatem, A. J., Holmes, E. C., & Bourhy, H. (2010). Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS pathogens*, **6**(10); page e1001,166. ISSN 1553-7374.

Tao, X. Y., Tang, Q., Li, H., Mo, Z. J., Zhang, H., Wang, D. M., Zhang, Q., Song, M., Velasco-Villa, A., Wu, X., Rupprecht, C. E., & Liang, G.-D. (2009). Molecular epidemiology of rabies in Southern People's Republic of China. *Emerging infectious diseases*, **15**(8); pages 1192–8. ISSN 1080-6059.

Tenzin, Sharma, B., Dhand, N., Timsina, N., & Ward, M. (2010). Reemergence of Rabies in Chhukha District, Bhutan, 2008. *Emerging infectious diseases*, **16**(12); pages 1925–2015.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**; pages 411–423. ISSN 1369-7412.

Tordo, N., Poch, O., Ermine, a., Keith, G., & Rougeon, F. (1988). Completion of the rabies virus genome sequence determination: highly conserved domains among the L (polymerase) proteins of unsegmented negative-strand RNA viruses. *Virology*, **165**(2); pages 565–76. ISSN 0042-6822.

Torres, C., Lema, C., Gury Dohmen, F., Beltran, F., Novaro, L., Russo, S., Freire, M. C., Velasco-Villa, A., Mbayed, V. A., & Cisterna, D. M. (2014). Phylodynamics of vampire bat-transmitted rabies in Argentina. *Molecular Ecology*, **23**(9); pages 2340–2352. ISSN 1365294X.

Townsend, S. E., Sumantra, I. P., Pudjiatmoko, Bagus, G. N., Brum, E., Cleaveland, S., Crafter, S., Dewi, A. P. M., Dharma, D. M. N., Dushoff, J., Girardi, J., Gunata, I. K., Hiby, E. F., Kalalo, C., Knobel, D. L., Mardiana, I. W., Putra, A. A. G., Schoonman, L., Scott-Orr, H., Shand, M., Sukanadi, I. W., Suseno, P. P., Haydon, D. T., & Hampson, K. (2013). Designing programs for eliminating canine rabies from islands: Bali, Indonesia as a case study. *PLoS neglected tropical diseases*, **7**(8); page e2372. ISSN 1935-2735.

Trewby, H., Wright, D., Breadon, E. L., Lycett, S. J., Mallon, T. R., Mccormick, C., Johnson, P., Orton, R. J., Allen, A. R., Galbraith, J., Herzyk, P., Skuce, R. A., Biek, R., & Kao, R. R. (2016). Use of bacterial whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis. *Epidemics*, **14**; pages 26–35. ISSN 1755-4365.

Tuite, A. R., Tien, J., Eisenberg, M., Earn, D. J. D., Ma, J., & Fisman, D. N. (2011). Cholera epidemic in Haiti, 2010: Using a transmission model to explain spatial spread of disease and identify optimal control interventions. *Annals of Internal Medicine*, **154**(9); pages 593–601. ISSN 00034819.

Turner, M. G. (1989). Landscape Ecology: The Effect of Pattern on Process. *Annual Review of Ecology and Systematics*, **20**(1); pages 171–197. ISSN 0066-4162.

van Etten, J. (2015). R Package gdistance: Distances and Routes on Geographical Grids.

VanDerWal, J., Falconi, L., Januchowski, Stephanie Shoo, L., & Storlie, C. (2014). SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises.

Velasco-Villa, A., Reeder, S. A., Orciari, L. A., Yager, P. A., Franka, R., Blanton, J. D., Zuckero, L., Hunt, P., Oertli, E. H., Robinson, L. E., & Rupprecht, C. E. (2008). Enzootic rabies elimination from dogs and reemergence in wild terrestrial carnivores, United States. *Emerging infectious diseases*, **14**(12); pages 1849–54. ISSN 1080-6059.

Viana, M., Cleaveland, S., Matthiopoulos, J., Halliday, J., Packer, C., Craft, M. E., Hampson, K., Czupryna, A., Dobson, A. P., Dubovi, E. J., Ernest, E., Fyumagwa, R., Hoare, R., Hopcraft, J. G. C., Horton, D. L., Kaare, M. T., Kanellos, T., Lankester, F., Mentzel, C., Mlengeya, T., Mzimbiri, I., Takahashi, E., Willett, B., Haydon, D. T., & Lembo, T. (2015). Dynamics of a morbillivirus at the domesticâĂŞwildlife interface: Canine distemper virus in domestic dogs and lions. *Proceedings of the National Academy of Sciences*, **112**; pages 1464–1469. ISSN 0027-8424.

Vigilato, M. A. N., Clavijo, A., Knobl, T., Silva, H. M. T., Cosivi, O., Schneider, M. C., Leanes, L. F., Belotto, A. J., & Espinal, M. A. (2013). Progress towards eliminating canine rabies: policies and perspectives from Latin America and the Caribbean. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **368**(1623); page 20120,143. ISSN 1471-2970.

Walsh, M. (2008). Pastoralism and Policy Processes in Tanzania: Mbarali Case Study. Technical Report September, University of Cambridge.

Wandeler, A. I., Capt, S., Gerber, H., Kappeler, A., & Kipfer, R. (1988). Rabies epidemiology, natural barriers and fox vaccination. *Parassitologia*, **30**(1); pages 53–57. ISSN 00482951.

Wang, T. H., Donaldson, Y. K., Brettle, R. P., Bell, J. E., & Simmonds, P. (2001). Identification of shared populations of human immunodeficiency Virus Type 1 infecting microglia and tissue macrophages outside the central nervous system. *Journal of Virology*, **75**; pages 11,686–11,699.

Weiss, B., Hoffmann, U., Freuling, C., Müller, T., Fesseler, M., & Renner, C. (2009). Rabies exposure due to an illegally imported dog in Germany. *Rabies Bulletin Europe*, **33**; pages 5–7.

Weiss, R. a. & McMichael, A. J. (2004). Social and environmental risk factors in the emergence of infectious diseases. *Nature medicine*, **10**(12 Suppl); pages S70–6. ISSN 1078-8956.

Wheeler, D. C. & Waller, L. a. (2008). Mountains, valleys, and rivers: The transmission of raccoon rabies over a heterogeneous landscape.

WHO (2005). WHO Expert Consultation on Rabies. *World Health Organization technical report series*, **931**; pages 1–88.

WHO (2013). WHO Expert Consultation on Rabies. Second report. *World Health Organization technical report series*, (982). ISSN 05123054.

Wilson, M. L., Bretsky, P. M., Cooper G.H., J., Egbertson, S. H., Van Kruiningen, H. J., & Cartter, M. L. (1997). Emergence of raccoon rabies in connecticut, 1991-1994: Spatial and temporal characteristics of animal infection and human contact. *American Journal of Tropical Medicine and Hygiene*, **57**(4); pages 457–463. ISSN 00029637.

Windiyaningsih, C., Wilde, H., Meslin, F. X., Suroso, T., & Widarso, H. S. (2004). The Rabies Epidemic on Flores Island , Indonesia (1998-2003). *Journal of the Medical Association of Thailand = Chotmaihet thangphaet*, **87**(11); pages 1389–93. ISSN 0125-2208.

Woodroffe, R. & Donnelly, C. a. (2011). Risk of contact between endangered African wild dogs Lycaon pictus and domestic dogs: opportunities for pathogen transmission. *Journal of Applied Ecology*, **48**(6); pages 1345–1354. ISSN 00218901.

Woolhouse, M., Dye, C., Etard, J., Smith, T., Charlwood, J., Garnett, G., Hagan, P., Hii, J., Ndhlovu, P., Quinnell, R., & Others (1997). Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proceedings of the National Academy of Sciences*, **94**(1); page 338.

Woolhouse, M. E. J. (2002). Population biology of emerging and re-emerging pathogens. *Trends in microbiology*, **10**(10 Suppl); pages S3–7. ISSN 0966-842X.

Wu, J. (2004). Effects of changing scale on landscape pattern analysis: Scaling relations. *Landscape Ecology*, **19**(2); pages 125–138. ISSN 09212973.

Ypma, R. J. F., Bataille, a. M. a., Stegeman, A., Koch, G., Wallinga, J., & van Ballegooijen, W. M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings. Biological sciences / The Royal Society*, **279**(1728); pages 444–50. ISSN 1471-2954.

Ypma, R. J. F., Jonges, M., Bataille, A., Stegeman, A., Koch, G., van Boven, M., Koopmans, M., van Ballegooijen, W. M., & Wallinga, J. (2013). Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *The Journal of infectious diseases*, **207**(5); pages 730–5. ISSN 1537-6613.

Zanoni, R. & Breitenmoser, U. (2003). Rabies in a puppy in Nyon, Switzerland. *Rabies Bulletin Europe*, **27**; pages 5–7.

Zeller, K. a., McGarigal, K., & Whiteley, A. R. (2012). Estimating landscape resistance to movement: A review. *Landscape Ecology*, **27**(6); pages 777–797. ISSN 09212973.